

Designing and Validating an Evidence-Based Pedagogy Framework: A PRISMA–Delphi–IRR Approach for Technical Education

Arthur C. Dela Peña¹

Received: November 11, 2025 – Revised: December 26, 2025 – Accepted: January 8, 2026

Abstract

Instructional quality in aviation maintenance training (AMT) is commonly regulated through content requirements, with limited attention to pedagogy and instructional assurance. This study develops and validates RAMP-ACE, an auditable framework for improving the reliability and transparency of competency-based teaching. A PRISMA-guided systematic review initially identified seven preliminary pedagogical domains. Through regulatory triangulation and Delphi refinement, overlapping standardization/calibration and quality-assurance constructs were consolidated, producing a final six-domain framework. The candidate item pool progressed from 14 indicators in v1 to 23 indicators in v2, with 21 retained after two indicators failed Delphi thresholds. Content validity was examined through a two-round asynchronous Delphi process (Round 1 $N = 14$; Round 2 $N = 13$), followed by a pilot audit assessing interrater reliability, rating time, artifact availability, and corrective-action closure. Delphi findings showed strong consensus, with most indicators meeting I-CVI ≥ 0.78 , S-CVI/Ave = 0.89, and S-CVI/UA = 0.35. Pilot testing produced κ /ICC values of 0.72–0.86, median dossier rating time of 38 minutes, and artifact availability above 85% for key records. Corrective actions were mostly closed within two weeks, supporting operational feasibility and transferability to technical and competency-based education contexts.

Keywords: Research Methodology; Instrument Development; Competency-Based Education; Assessment Rubric; Quality Assurance; Technical and Vocational Education and Training (TVET); Delphi Method

¹ Corresponding Author,

National Aviation Academy of the Philippines, Pampanga, Philippines

Email: artair248@gmail.com

Introduction

Aircraft maintenance is a safety-critical domain where training quality directly shapes operational reliability and risk management. Aviation Maintenance Technicians (AMTs) operate in high-risk environments where seemingly minor errors—such as torque misapplication, flawed troubleshooting, or tool control lapses—can escalate into serious incidents (Karunakaran & Ashok Babu, 2022; Franciosi et al., 2019). Instructional quality is therefore not just an educational issue but a cornerstone of safety assurance. Over the past decade, training systems have increasingly adopted competency-based training and assessment (CBTA), which prioritizes measurable performance outcomes over time-based instruction (Yusop et al., 2022; Chen et al., 2022). This approach requires instructors to demonstrate pedagogical competence, including outcome-based design, assessment literacy, effective feedback strategies, and integration of human factors. When pedagogy is assumed rather than developed, training quality becomes inconsistent, assessment validity weakens, and standardization is lost (James et al., 2020; Chen et al., 2024).

Global regulatory frameworks have started to address this issue explicitly. The International Civil Aviation Organization (ICAO) defines instructor and evaluator roles through its training standards and TRAINAIR PLUS Instructor Course (ICAO, 2020, 2025). The Federal Aviation Administration (FAA) integrates instructional expectations into its instructor handbook, while the European Union Aviation Safety Agency (EASA) embeds teaching methods into Part-147 compliance frameworks (FAA, 2022; EASA, 2023). The International Air Transport Association (IATA) and International Federation of Air Line Pilots' Associations (IFALPA) similarly link instructor competence to CBTA performance monitoring (IATA & IFALPA, 2021; IATA, 2025). Collectively, these frameworks treat pedagogical competence as an auditable safety variable.

In contrast, the Civil Aviation Authority of the Philippines PCAR Part 3 focuses on technical qualifications and curricula but does not codify pedagogical competence for AMT instructors (Clare & Kourousis, 2021; Thulasy et al., 2022). This regulatory gap becomes more critical as training evolves to include emerging technologies such as advanced composites, digital troubleshooting, and electronic logbooks (Vlasblom et al., 2020; Tyagi et al., 2023). Without explicit pedagogical standards, instructional quality remains uneven, and regulators lack robust audit criteria.

This study addresses this gap by developing and validating the Regulatory-Aligned Maintenance Pedagogy-Assessment, Calibration & Evidence (RAMP-ACE) framework. RAMP-ACE makes instructional competence explicit, measurable, and auditable within ATO quality systems by defining domains, indicators, proficiency

levels, and evidence types aligned with international standards (ICAO, 2020, 2025; FAA, 2022; EASA, 2023). The framework is built through an integrated methodological architecture that combines comparative regulatory analysis, a systematic literature review, and Delphi-based expert consensus to ensure rigor, alignment, and practical feasibility (Mitchell, 2025; Yusop et al., 2023). Although demonstrated in aviation maintenance training, the validation model is designed to be transferable to other TVET settings that require structured, evidence-based competency assurance.

The study is guided by three questions: Which pedagogical domains are consistently recognized across international and TVET standards relevant to AMT instruction? How do PCAR Part 3 provisions diverge from these domains? And what consensus-based indicators and evidence are necessary to make instructional competence auditable in the Philippine context? Addressing these questions produces a regulatory-ready pedagogical framework, a clause-to-domain crosswalk between PCAR and global regimes, and a methodological model that can be adapted by other technical training systems and regulatory bodies (Hovlid et al., 2020; Piggott et al., 2021).

The subsequent sections present the literature foundation, methodological processes, and key results from the regulatory analysis, SLR, Delphi validation, and pilot audit, followed by policy and implementation guidance for integrating pedagogical competence into PCAR Part 3. Through this structured and transferable approach, RAMP-ACE provides a scalable model for standardized AMT instruction and contributes to broader TVET and safety-critical training oversight.

Literature Review

1. AMT Training, Human Factors, and Error Pathways

Aircraft maintenance remains vulnerable to human factors, with error pathways creating latent risks that may surface long after a task is completed. Systematic analyses confirm the strong contribution of human factors to maintenance-related safety outcomes (Franciosi et al., 2019; Karunakaran & Ashok Babu, 2022). Proactive, predictive risk-management approaches supported by artificial intelligence and strategic risk mapping have been emphasized to mitigate these hazards (Mendes et al., 2022; Küçük Yılmaz, 2019). Gaps remain in existing training systems, particularly in Europe, where limited guidance on incident learning and the absence of standardized root-cause analysis frameworks restrict the effectiveness of human factors instruction (Clare & Kourousis, 2021). High-quality training that embeds human factors is therefore indispensable for addressing latent failures and strengthening systemic resilience. However, empirical research directly linking AMT pedagogy to

reductions in error remains scarce, with few standardized competency frameworks and validated assessment tools designed for maintenance work.

Within broader educational research, structured competency models and validated teaching frameworks—including Danielson’s Framework for Teaching, the Marzano Teacher Evaluation Model, and Technological Pedagogical Content Knowledge (TPACK)—have been used to operationalize teacher quality via domains, indicators, and performance levels, and to examine validity and links to student outcomes (Basileo & Toth, 2019; Jiménez Sierra et al., 2023; Morris-Mathews et al., 2021). Positioning AMT training within this broader discourse underscores the need for similarly auditable, evidence-based pedagogical frameworks that support both safety and learning outcomes, rather than relying solely on regulatory content prescriptions.

2. Training Effectiveness Across Safety-Critical Domains

Evidence from safety-critical industries reinforces the value of structured, competency-based instruction. Proficiency-based progression has reduced procedural errors and training time compared with time-based models (Mazzone et al., 2020), while virtual reality has improved knowledge acquisition and short-term retention, even though long-term effects remain underexplored (Scorgie et al., 2024). In healthcare, simulation-based training enhances teamwork and human factors skills, and realistic human simulation can reduce medication errors when paired with expert facilitation (Abildgren et al., 2022; Sarfati et al., 2019). However, assessment methods in these settings vary widely, and standardized evaluation tools are often lacking, particularly for vulnerable or underrepresented worker groups (Peiró et al., 2020).

Across these domains, training innovations frequently outpace the development of robust, validated frameworks for judging instructional quality and competence. Evaluation is still commonly driven by local checklists or ad hoc rating criteria rather than auditable, theory-informed models. This mirrors challenges in AMT, where advanced instructional technologies and human factors content are increasingly used, but are not systematically embedded in a competency framework with clear domains, indicators, and evidence requirements. Situating AMT within this wider pattern underscores the need for a structured, validated, and audit-ready pedagogical framework such as RAMP-ACE.

3. Competency-Based Training in Technical and Vocational Education

Competency-based training and assessment (CBT/CBA) has emerged as the dominant model across technical and vocational education. In medicine, competency-based medical education has restructured curricula and assessment around clearly defined outcomes (Mitchell, 2025). Systematic reviews in TVET confirm

the effectiveness of competency-based assessment in enhancing skill development and learning outcomes (Yusop et al., 2022). Nursing education also reports significant improvements in clinical skills, critical thinking, and professional growth when CBT is implemented effectively (Chen et al., 2022). Successful implementation requires integrating technical and twenty-first-century skills, such as adaptability and problem-solving, aligned with Industry 4.0 (Yusop et al., 2023). Persistent challenges include inadequate facilities, insufficient instructor preparation, and weak industry engagement, though partnerships between schools and employers have proven beneficial (Suharno et al., 2020). Despite the clear benefits of CBT/CBA, AMT training remains underexplored, with few validated AMT-specific assessment tools, limited Industry 4.0 integration, and minimal research on faculty development.

4. Instructor Competence and International Gaps

Although global frameworks increasingly treat instructor competence as observable and auditable, significant deficiencies persist in aviation maintenance. European regulations, for example, lack explicit requirements for human factors trainers, contributing to inconsistent instructional quality (Clare & Kourousis, 2021). Maritime education faces similar challenges: while the STCW Convention mandates qualified instructors, it offers limited guidance on systematic pedagogical development (Vujičić et al., 2020). Across safety-critical domains, rapid digitalization and evolving safety communication demands require new competencies in instructional design, contextualization, and learner engagement (Thulasy et al., 2022; Tyagi et al., 2023). Research also shows that minimalist, user-centered instructional designs can improve comprehension and reduce cognitive overload, particularly for complex procedures (Heinonen et al., 2022).

These findings underscore the need for standardized AMT instructor frameworks that integrate technical expertise with validated pedagogical competence. In contrast, general education and teacher education increasingly rely on structured competency and evaluation frameworks—such as Danielson-type teaching frameworks, Marzano-style models, and TPACK-derived approaches—that define domains, performance levels, and evidence for evaluating instructional quality. However, no universal, auditable pedagogical model exists for AMT, and empirical evidence linking instructor pedagogy to safety outcomes remains limited. This methodological gap motivates the development of an auditable, framework-based approach to AMT instructor competence, such as RAMP-ACE, that can align maintenance training with contemporary educational quality assurance and accreditation practices.

5. Migration of Standards and Regulatory Frameworks

Literature on regulatory theory explains how professional standards evolve across sectors through structured change processes. In higher education, institutions adapt to accreditation pressures by using change management frameworks to ensure compliance with external quality standards (Edge et al., 2021; Paquibut & Al Naamany, 2020). In healthcare, inspections help organizations identify performance gaps and align practices with regulatory expectations (Hovlid et al., 2020). Neo-institutional theory highlights how coercive, mimetic, and normative pressures drive harmonization across jurisdictions (Haapamäki, 2022). Historical analyses of professions like midwifery and nursing illustrate how demographic and policy shifts fostered outcome-based regulatory models (Safari et al., 2023; Stievano et al., 2019). Aviation faces similar pressures as domestic regulators adopt international standards, often encountering heterogeneity and redundancy (Edgar & Stack, 2023). These patterns demonstrate the feasibility—and necessity—of integrating international pedagogical standards into PCAR Part 3. However, tested models remain scarce, particularly regarding effective enforcement mechanisms, performance indicators, and comparative policy strategies for regulatory alignment. To further illustrate cross-sector validation practices and support methodological transferability, Table 1 compares how different fields validate teaching or competency frameworks and situates the present study's approach within this landscape.

Table 1

Validation Approaches for Teaching and Competency Frameworks Across Fields

Field	Common Validation Basis	Typical Evidence	Key Limitation	Contribution of RAMP-ACE
General Education	Expert consensus (e.g., Danielson, Marzano)	Observations, rubrics	Limited auditability; scorer subjectivity	Adds auditable artifacts and IRR testing
EdTech / Teacher Education	Construct validation (e.g., TPACK)	Surveys, factor models	Reliance on self-report data	Incorporates document-based verification
Healthcare Education	Competency-based, simulation-focused	Checklists, expert ratings	Inconsistent calibration across raters	Formalizes calibration and reliability thresholds
Maritime Training	Regulatory compliance checks	Licenses, curriculum audits	Pedagogy weakly operationalized	Translates pedagogy into observable indicators
TVET	Outcomes-based standards	Skills checklists, assessments	Fragmented QA mechanisms	Integrates pedagogy, assessment, and QA
This study (AMT)	PRISMA–Delphi–IRR multi-source validation	Literature, expert consensus, audit artifacts	Addresses construct ambiguity and scoring reliability	Transferable, audit-ready validation architecture

Conceptual Model

The study proposes the Regulatory-Aligned Maintenance Pedagogy-Assessment, Calibration & Evidence (RAMP-ACE) framework, which operationalizes pedagogical competence for AMT instructors by translating international regulatory guidance into a PCAR Part 3-compatible instructional quality-assurance model. RAMP-ACE responds to evidence that maintenance risk decreases when training systems move from reactive, experience-based instruction to proactive, evidence-driven pedagogy (Clare & Kourousis, 2021; Mendes et al., 2022). To establish theoretical grounding, the framework is explicitly aligned with constructive alignment principles (coherence across outcomes, instruction, and assessment) and reflects the evaluative logic of CIPP (Process–Product) and Kirkpatrick’s learning–behavior–results hierarchy. The primary unit of analysis is the instructor dossier, which allows indicators to scale to the course and organizational levels.

RAMP-ACE integrates three evidence streams—comparative regulatory mapping, systematic literature review (SLR) findings, and Delphi expert judgment—to derive its domain structure. Although the systematic literature review initially identified seven pedagogical domains, triangulation with regulatory mapping and Delphi consensus indicated substantial conceptual overlap between quality assurance (QA) and standardization constructs; these were therefore integrated into a single domain during framework consolidation, resulting in six final RAMP-ACE domains. Table 2 summarizes how ICAO, FAA, EASA, IATA–IFALPA, and PCAR Part 3 converge on lesson planning, safety pedagogy, assessment validity, and instructor standardization, while also revealing gaps (e.g., human-performance pedagogy, structured reflection). These aligned and divergent elements, reinforced by SLR themes and panel consensus, justify selecting six domains as the minimal, recurrent set of observable pedagogical behaviors supported across both global regulatory guidance and established educational competency models.

RAMP-ACE is organized across five interdependent layers (Figure 1):

Layer 1: Alignment Engine (Regulatory Layer). PCAR Part 3 clauses were mapped against ICAO Doc 9868/TIC, FAA’s Aviation Instructor’s Handbook, EASA Part-147 AMC/GM, and IATA–IFALPA guidance. Each clause was tagged as must/should/may, producing a crosswalk that identifies regulatory gaps and specifies the scope of auditable instructional requirements.

Layer 2: Competency Model (Pedagogy Layer). Six domains emerged from triangulation across the SLR, regulatory mapping, and Delphi results: (1) CBTA-aligned lesson design, (2) lab and safety pedagogy, (3) assessment literacy, (4) human-performance learning, (5) facilitation and inclusion, and (6) standardization and reflective practice. Each domain includes behavioral indicators linked to verifiable evidence (e.g., lesson plans, assessment blueprints, hazard briefs).

Layer 3: ACE Cycle (Assurance Layer). A structured loop of Assessment → Calibration → Enhancement maintains instructional reliability through rubric scoring, peer norming, and corrective-action tracking, mirroring contemporary educational quality-assurance models emphasizing data-driven improvement.

Layer 4: Evidence Ledger (Proof Layer). Curated artifacts—lesson plans, assessment blueprints, safety briefs, observation notes, and calibration records—serve as standardized, auditable evidence for internal quality assurance and external oversight.

Layer 5: Maturity Model (Growth Layer). Instructor development is benchmarked across five levels (Absent → Exemplary) using metrics such as indicator coverage, calibration health, and corrective-action closure.

Collectively, these layers transform instructional quality from tacit practitioner knowledge into a theoretically grounded, evidence-based system suitable for regulatory oversight and broader educational evaluation. The revisions clarify the conceptual grounding, integration logic, and unit of analysis, directly addressing the reviewer's concerns about methodological transparency and transferability to TVET and broader education research.

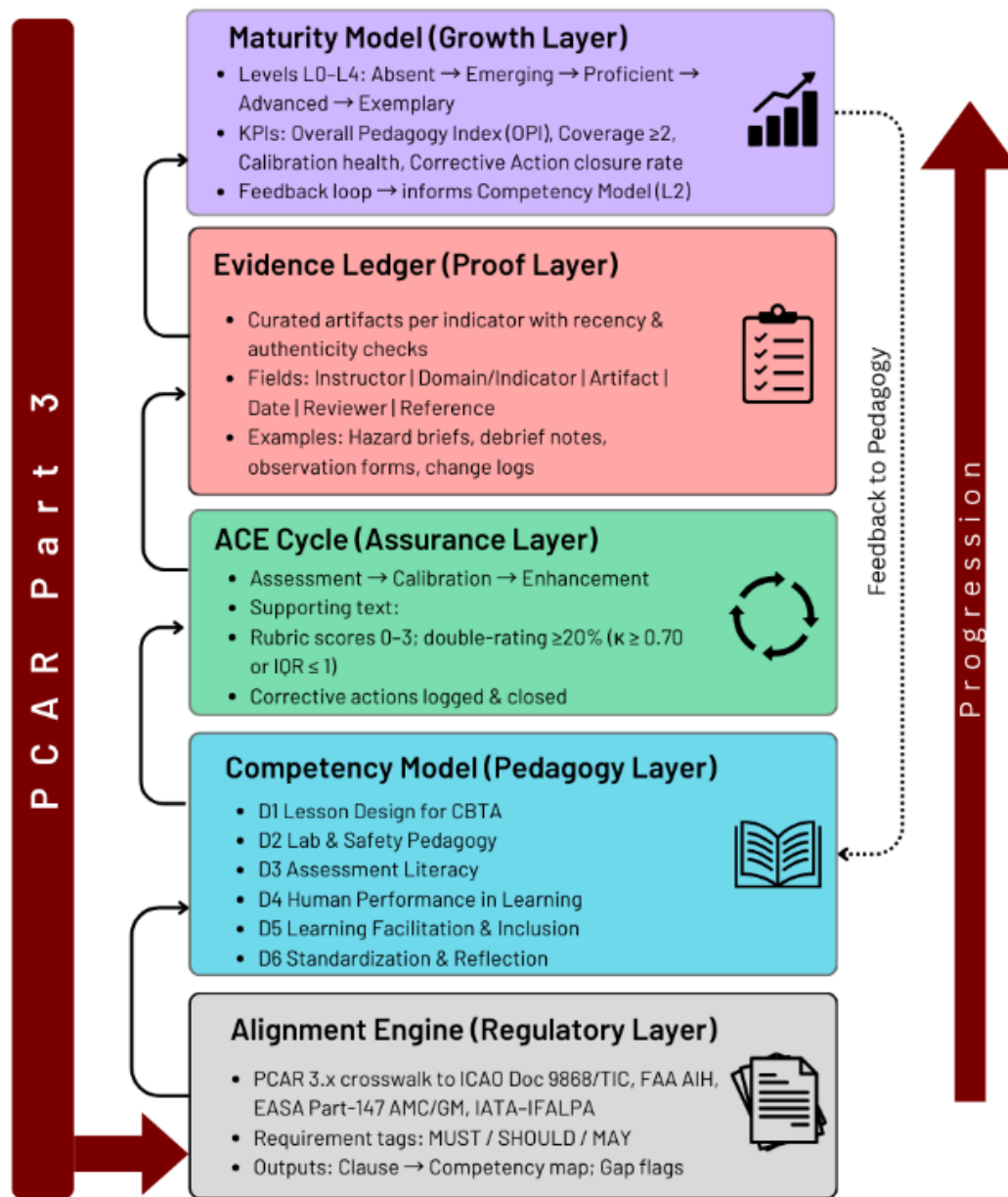
Table 2

Cross-Regime Mapping of Instructor Competence Requirements and Pedagogical Expectations.

Competency Area	PCAR Part 3	ICAO Doc 9868/TIC	FAA AIH	EASA Part-147 AMC/GM	IATA-IFALPA	Implication for RAMP-ACE Domain
Lesson Design & Outcomes Alignment	Limited guidance	Strong emphasis on structured competency-based design	High emphasis on instructional planning	Requires documented training syllabi	Emphasizes structured lesson planning	Supports D1 Lesson Design for CBTA
Safety & Lab Pedagogy	Weak/implicit	Embedded in HP and risk-based training	Includes safety briefings	Requires safety oversight in training	Emphasizes operational risk awareness	Justifies D2 Lab & Safety Pedagogy
Assessment Literacy & Validity	Minimal specificity	Strong in competency frameworks	Detailed methods & validity principles	Requires assessment procedures	Reinforces performance-based assessment	Supports D3 Assessment Literacy
Human Performance & Facilitation	Not explicitly required	Core HP model	Covers motivation & learning theory	Limited	Focus on human factors	Justifies D4 Human Performance in Learning
Inclusive Learning & Facilitation	Not covered	Addresses facilitation methods	Strong focus on learner-centered methods	Not explicit	Encourages contextual adaptation	Supports D5 Facilitation & Inclusion
Standardization & QA	Not required	Requires continuous improvement	Promotes reflective practice	Requires oversight system	Encourages instructor standardization	Leads to D6 Standardization & Reflection

Figure 1

RAMP-ACE Conceptual Model



Methods

1. Design

This study employed an integrated, multi-source validation design that combined regulatory document analysis, systematic evidence synthesis, expert consensus, and empirical reliability testing. The approach is consistent with mixed-methods triangulation and established document-analysis traditions in educational research, where converging evidence from multiple sources strengthens construct clarity, interpretive robustness, and validation credibility. Regulatory texts (e.g., PCAR Part 3, ICAO Doc 9868, FAA Aviation Instructor’s Handbook, EASA Part-147 AMC/GM,

and IATA–IFALPA guidance) were systematically coded into domains, indicators, and auditable evidence types and then consolidated with themes derived from the literature to generate an initial competency architecture.

A two-round asynchronous Delphi process, aligned with contemporary consensus-reporting standards (Niederberger et al., 2024), was conducted to assess indicator clarity, relevance, and necessity using median ratings and interquartile range (IQR) thresholds. Experts were purposively selected based on predefined criteria—specifically, a minimum of five years of experience in ATO operations, regulatory oversight, CBTA implementation, or human-factors instruction—to ensure multidomain and practice-relevant representation.

A pilot audit constituted the fourth validation layer, evaluating feasibility and scoring reliability through instructor dossiers assessed using the RAMP-ACE rubric. Interrater agreement ($\kappa \geq 0.70$) and artifact availability were used as indicators of implementation readiness (Teresi et al., 2022). Unlike Delphi-only or single-source validation designs that rely primarily on expert judgment at a single point in time, this multi-source validation architecture addresses known limitations related to construct under-specification, interpretive ambiguity, and untested scoring reliability by integrating documentary evidence synthesis, structured expert consensus, and empirical interrater testing within a single, audit-ready framework.

This integrated model advances prevailing validation practices in educational research by operationalizing constructive alignment and CIPP-inspired evaluation logic in a form that is both methodologically rigorous and regulator-ready. Although demonstrated in an aviation maintenance training context, the design is transferable to TVET and other regulated training environments that require evidence-based competency frameworks and standardized assessment assurance.

2. Data Sources

This study draws on three complementary data sources to ensure regulatory alignment, empirical grounding, and operational relevance.

First, a regulatory corpus was assembled comprising authoritative aviation training and oversight instruments, including PCAR Part 3, ICAO Doc 9868 (PANS-TRG), TRAINAIR PLUS Instructor Course materials, the FAA Aviation Instructor's Handbook, EASA Part-147 AMC/GM, and IATA–IFALPA instructor guidance. These documents established the normative baseline for defining instructor competence and identifying auditable pedagogical expectations across jurisdictions.

Second, an empirical literature corpus was constructed from peer-reviewed studies on competency-based training, instructional quality assurance, assessment validity, and instructor standardization in technical and safety-critical education.

This corpus informed the theoretical grounding of the framework and the identification of recurring pedagogical domains and indicators. The procedures used to identify, screen, and synthesize this literature are reported separately in the Systematic Literature Review subsection.

Third, an institutional artifact set was examined to assess operational feasibility. This set included anonymized training manuals, lesson and assessment templates, observation instruments, standardization records, and internal QA reports from Approved Training Organizations. These artifacts were used to evaluate evidence availability, rating effort, and implementation feasibility under real training conditions, rather than to generate new indicators. Together, these sources enabled triangulation among what regulations require, what the literature supports, and what institutions can realistically evidence, providing a robust foundation for developing an auditable, implementation-ready pedagogical framework.

3. Comparative Regulatory Analysis

A clause-level comparative regulatory analysis was conducted to examine how PCAR Part 3 aligns with major international training and oversight frameworks, including ICAO Doc 9868/TIC, FAA AIH/Part 147, EASA Part-147 AMC/GM, and IATA–IFALPA guidance. The analysis used a deductive coding frame composed of seven regulatory coding dimensions: instructor qualification, pedagogy, CBTA delivery, assessment literacy, recurrent upskilling, evidence-based instructional quality assurance, and evidence requirements. These seven dimensions served only as analytical categories for comparing regulatory provisions and should not be interpreted as the final RAMP-ACE domain structure. Inductive coding was also applied to capture context-specific provisions not explicitly represented in the international standards.

All regulatory texts were dual-coded at the smallest enforceable clause or provision level. Disagreements were resolved through adjudication after a calibration pass. Methodological rigor was supported through reliability checks, including Cohen's κ for dimensional presence or absence, with a target threshold of ≥ 0.70 , and Krippendorff's α for ordinal strength ratings, with a target threshold of ≥ 0.67 . Each clause was further assigned an evidence-based instructional quality-assurance level (A0–A3) and a requirement-intensity rating (MUST/SHOULD/MAY) to support policy translation and auditability.

The resulting Alignment Engine generated a consolidated heatmap and clause-to-competency crosswalk that identified regulatory gaps, areas of international convergence, and auditable evidence expectations. These outputs informed, but did not directly determine, the RAMP-ACE competency architecture. The final RAMP-ACE

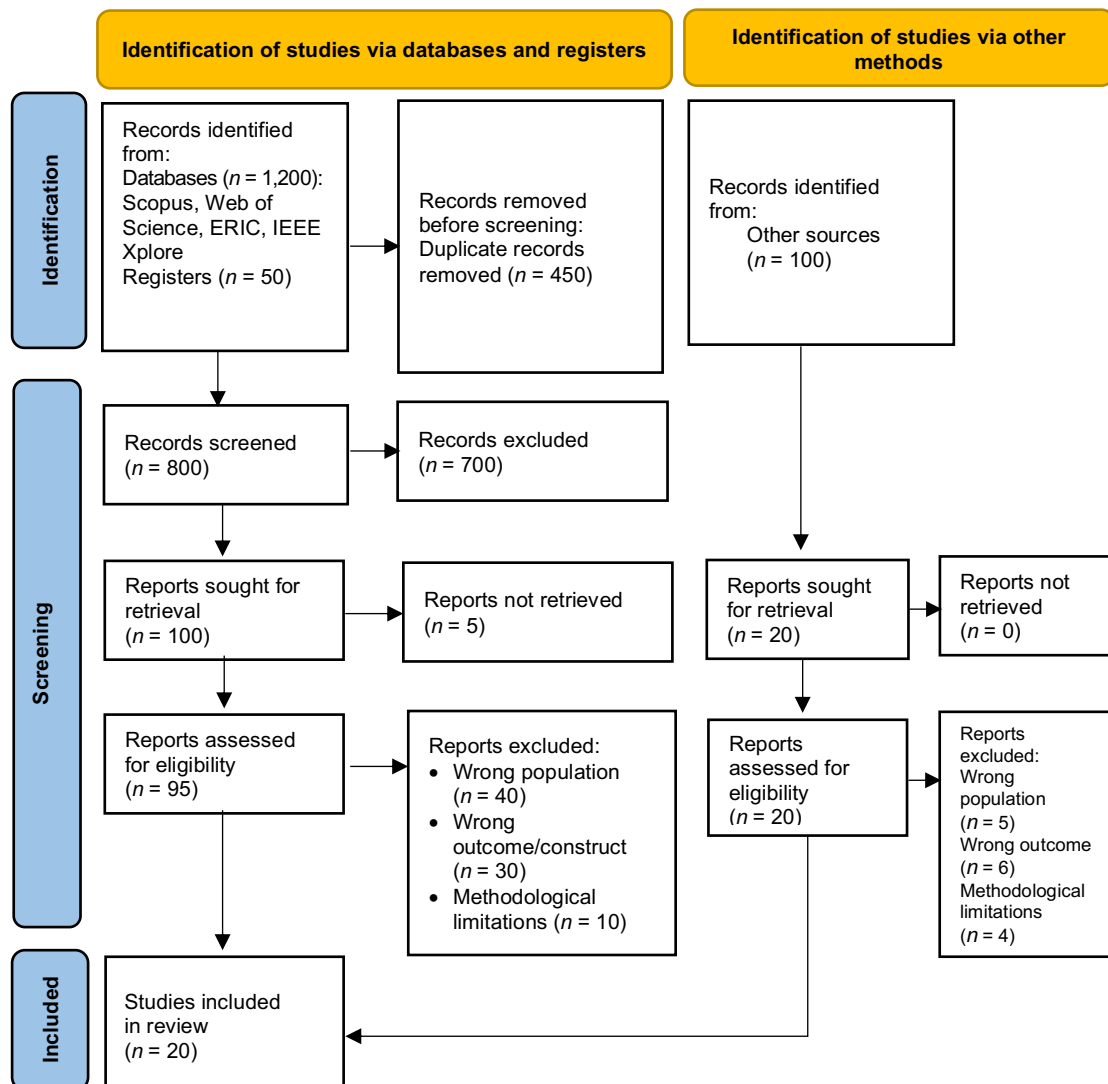
framework was subsequently derived through triangulation of the regulatory analysis, SLR evidence, and Delphi validation, resulting in the validated six-domain structure reported in the final framework.

4. Systematic Literature Review

A systematic literature review was conducted following PRISMA 2020 guidelines to identify evidence on instructor competence, CBTA, assessment validity, instructional standardization, and safety-critical technical education. Searches were conducted in Scopus, Web of Science, ERIC, and IEEE Xplore using database-adapted Boolean strings combining terms for competency-based training, instructors, assessment, instructional quality, TVET, technical education, safety-critical training, and aviation. Peer-reviewed studies with explicit methodologies were included, while non-journal sources were limited to authoritative regulatory or training-oversight guidance.

Figure 2

PRISMA 2020 flow diagram



As shown in Figure 2, 1,250 records were identified from databases and registers, comprising 1,200 database records and 50 register records. After 450 duplicates were removed, 800 records were screened, and 700 were excluded. Of the 100 reports sought for retrieval, five were not retrieved, and 95 were assessed for eligibility. Full-text exclusions were due to wrong population ($n = 40$), wrong outcome or construct ($n = 30$), and methodological limitations ($n = 10$), yielding 15 studies from databases/registers. Through other methods, 100 records were identified, 20 reports were sought and assessed for eligibility, and 15 were excluded because of wrong population ($n = 5$), wrong outcome ($n = 6$), or methodological limitations ($n = 4$), yielding five additional studies. Overall, 20 studies were included in the final synthesis. Two reviewers independently screened and extracted data using a calibrated template, followed by abductive–deductive thematic synthesis to identify recurring pedagogical domains, indicators, assessment practices, and types of auditable evidence. These findings informed the initial AMT-specific framework and supported the subsequent Delphi validation and framework refinement.

5. Framework Drafting (pre-Delphi)

Before expert validation, a rubric was drafted that operationalizes AMT instructor pedagogy into auditable units. Each entry follows a fixed schema—Domain → Indicator → Behavior descriptor → Acceptable evidence → Proficiency anchors (Emerging/Proficient/Advanced) → PCAR/International reference—so that practice can be observed, evidenced, and traced to regulation. Indicators are scored 0–3 (0 = not evident, 1 = partially evident, 2 = meets standard, 3 = exceeds standard with generalizable practice), with rater notes capturing the context and edge cases.

An Evidence Ledger links every indicator to verifiable artifacts (e.g., lesson plans, assessment blueprints, hazard briefs, peer-observation forms, debrief notes) and records provenance, reviewer initials, and clause references; a currency rule requires that artifacts be ≤ 24 months old (or aligned to the most recent course revision) unless a justified exemption is logged. At the domain level, anchors specify observable behaviors that differentiate Emerging versus Proficient and Advanced performance; at the framework level, an Overall Pedagogy Index (OPI) is computed as the mean of indicator scores across the six domains, enabling roll-up reporting, risk-based targeting for improvement, and Delphi-ready clarity about what constitutes “meeting” versus “exceeding” the standard.

6. Delphi Consensus

A two-round asynchronous Delphi process was conducted to refine and validate the v2 candidate framework using contemporary consensus-reporting standards. Experts were purposively selected based on predefined criteria, including at least five

years of experience in ATO quality assurance, AMT instruction, CBTA implementation, regulatory oversight, or human-factors training, with active involvement in curriculum, assessment, or standardization processes. The panel represented the main areas required to evaluate an auditable pedagogical competency framework and comprised specialists in regulatory oversight, AMT instruction, CBTA, and human factors (Round 1: $N = 14$; Round 2: $N = 13$).

Panelists rated each candidate indicator for clarity, relevance, and necessity. The analysis generated item medians, interquartile ranges, and item-level content validity indices. Indicators were retained when they met the predefined thresholds of median ≥ 4 , IQR ≤ 1 , and I-CVI ≥ 0.78 . Indicators falling below threshold were revised, removed, or reclassified based on panel feedback, regulatory alignment, and implementation feasibility. Consensus stability was assessed through median shifts across rounds, with changes of ≤ 0.5 interpreted as acceptable stability.

To avoid duplicating partial Delphi results, the complete decision trail is reported in the Results section. Table 5 presents the final candidate-to-framework mapping, including candidate codes, final RAMP-ACE codes, Round 2 medians, IQRs, I-CVI values, and final indicator decisions. This structure preserves traceability from the pre-final seven-domain candidate framework to the final validated six-domain RAMP-ACE framework.

7. Pilot Audit: Feasibility and Reliability

A document-based pilot audit was conducted to evaluate the feasibility of implementation and the reliability of scoring, using the instructor dossier as the primary unit of analysis. Dossiers—containing lesson plans, assessment blueprints, hazard briefs, observation notes, and standardization records—were scored with the RAMP-ACE rubric. Two trained raters independently evaluated all indicators, verified artifact provenance, and completed a calibration session to establish shared scoring anchors, consistent with performance-based competency assessment in educational evaluation research.

Interrater agreement was computed at the indicator level using ordinal agreement metrics. A random subset of dossiers was rescored to estimate reliability, yielding values above the predefined threshold (e.g., $\kappa = 0.74$; 95% CI [0.68, 0.82]), demonstrating stable, defensible scoring. Feasibility metrics included rating time, artifact completeness, and missing-data patterns. Operational artifacts, such as torque-card entries and blueprint-task alignment sheets, were critical for testing scoring consistency across indicators, such as “critical-step assurance,” and for illustrating how evidence items supported methodological validation rather than merely reflecting field procedures.

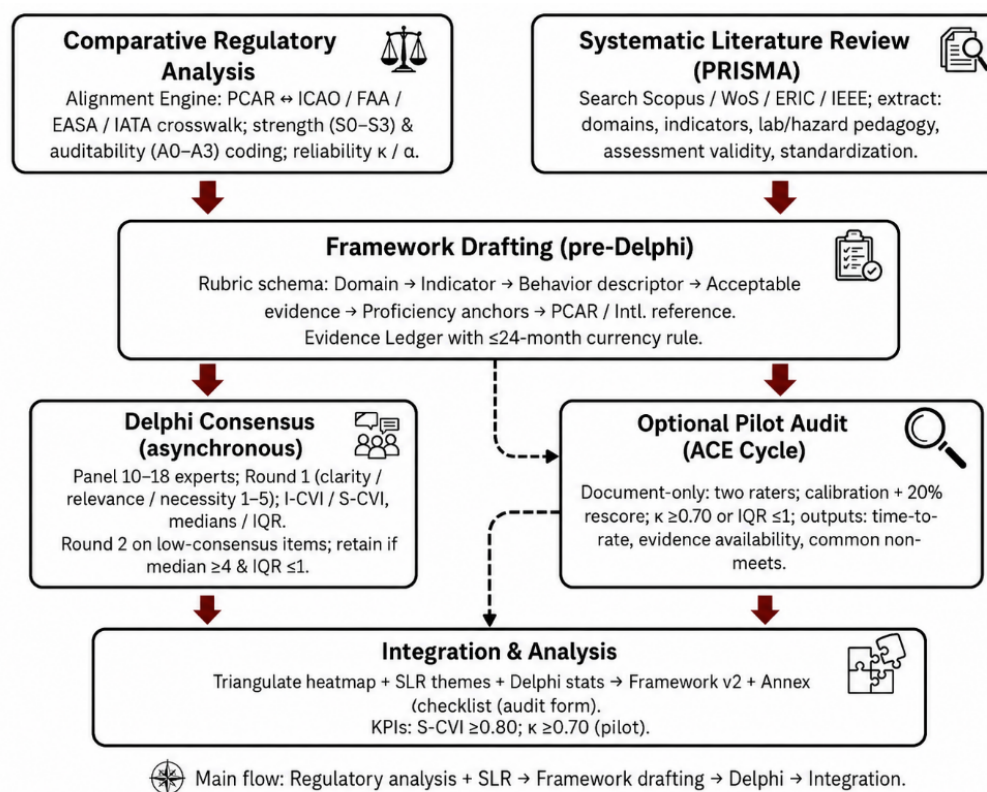
Audit outputs—evidence-availability heatmaps, exemplar non-meets, and corrective-action recommendations—fed directly into the ACE cycle. This approach mirrors CIPP “Process–Product” logic and aligns with Kirkpatrick’s emphasis on observable performance change, reinforcing the methodological contribution of RAMP-ACE. Pilot findings informed rubric refinement, calibration updates, and targeted instructional-improvement actions, demonstrating readiness for operational deployment and cross-institutional replication.

8. Integration & Analysis

Outputs were integrated through methodological triangulation, merging (i) the regulatory heatmap and crosswalk (coverage/strength/evidence-supported verification by clause), (ii) SLR thematic signals (competency domains, indicators, evidence types, calibration practices), and (iii) Delphi statistics (item medians, IQRs, I-CVI, S-CVI/Ave) to produce Framework v2 and an Annex checklist (audit form) (Figure 3). Discrepant cases—such as strong regulatory coverage but low Delphi agreement—were adjudicated in a structured review that prioritized evidence-based instructional quality assurance and implementation feasibility. Content validity was summarized at both item and scale levels. Items were retained, revised, or removed according to the Delphi decision rules and then mapped back to PCAR clauses to preserve traceability.

Figure 3

Study Design Flow (Convergent documentary + Delphi)



Key performance indicators (KPIs) guided acceptance: S-CVI/Ave ≥ 0.80 for the final item set; Cohen's $\kappa \geq 0.70$ (pilot) on presence/absence and domain IQR ≤ 1 on ordinal scores; percentage of indicators retained (target $\geq 70\%$ after one revision cycle); and qualitative panel-acceptance comments synthesized into actionable edits (e.g., clarified behavior descriptors, tighter evidence definitions). The Annex checklist operationalizes each indicator into auditable prompts and artifact fields (recency, source, reviewer), enabling direct uptake in ATO quality systems and alignment with the RAMP-ACE ACE cycle for ongoing calibration and enhancement.

9. Methodological Contribution and Transferability

This study introduces a multi-evidence validation pipeline that advances beyond conventional Delphi-only approaches commonly used in educational framework studies. By integrating PRISMA-guided content synthesis, Delphi expert consensus, and interrater reliability (IRR) testing through a pilot audit, the design produces convergent evidence of content validity, interpretive clarity, and scoring reliability. This triangulated structure operationalizes a comprehensive evaluation model consistent with contemporary educational measurement principles while remaining audit-ready for regulated training systems. Although developed in an aviation context, the validation sequence is transferable to broader TVET and professional education settings that require documented competency standards, calibrated assessment practices, and evidence-based quality assurance. The model also provides a replicable foundation for cross-institutional or cross-country adaptive validation.

Results

1. Alignment Engine Output

The clause-level crosswalk reveals a persistent regulatory gap in the conceptualization of pedagogical competence. While PCAR Part 3 specifies instructor qualification requirements (S2/A1), it provides minimal guidance on CBTA-aligned lesson design, assessment literacy, or the types of evidence needed to demonstrate instructional quality (typically S0–S1/A0–A1). In contrast, ICAO Doc 9868/TIC and IATA–IFALPA articulate explicit expectations for outcome-to-task alignment, criterion-referenced assessment, and performance evidence—core CBTA principles (S3/A3). FAA Part 147/AIH and EASA Part-147 AMC/GM similarly define auditable requirements across most pedagogical domains (S2/A2). Figure 4 visualizes these contrasts and shows how limited PCAR guidance leaves CBTA constructs operationally underdeveloped.

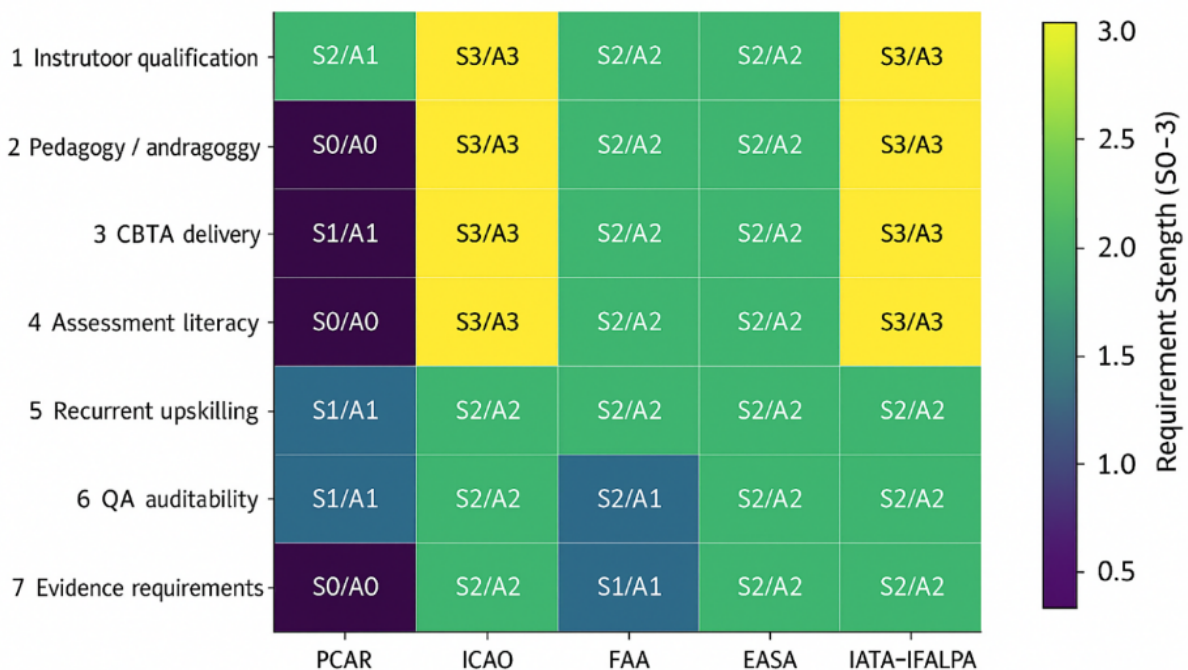
To ensure analytic defensibility, coding reliability met established thresholds: dual coding achieved Cohen's $\kappa \geq 0.70$ for presence/absence and Krippendorff's α

≥ 0.67 for ordinal strength ratings. Adjudication logs documented the reconciliation of discrepancies, and final strength and auditability ratings represent coder consensus. Evidence-based instructional QA tags (A0–A3) further differentiate regimes: PCAR rarely specifies required artifacts, whereas comparator frameworks identify concrete evidence (e.g., blueprints, scoring rubrics, debrief templates) aligned with CBTA’s emphasis on observable performance and documentation.

A brief coding example illustrates how regulatory insights shaped indicator development: clauses referencing “task analysis,” “performance criteria,” and “learning outcomes” across ICAO and FAA sources were consolidated under CBTA Lesson Design (D1), while statements on “calibration,” “standardized scoring,” and “rater agreement” informed Assessment Literacy (D3). These coded themes also aligned with the SLR findings, strengthening confidence that the identified gaps were theoretically substantive rather than regulatory omissions. Gaps cluster around three areas: (i) CBTA-based pedagogy and hazard-aware instructional delivery, (ii) assessment literacy and debrief evidence, and (iii) explicit artifact requirements and recency rules. These results directly informed the Annex checklist and RAMP-ACE domain structure, ensuring that proposed regulatory enhancements are internationally aligned, pedagogically grounded, and operationally auditable (Figure 4).

Figure 4

PCAR–International Crosswalk & documented quality assurance controls heatmap.



2. SLR synthesis: preliminary SLR-derived domains, indicators, and evidence types for AMT

The SLR identified seven preliminary pedagogical domains for AMT instruction: CBTA-aligned lesson design, laboratory and safety pedagogy, assessment literacy and blueprinting, standardization and calibration, human-performance learning, facilitation/inclusion and instructional communication, and quality assurance/continuous improvement. These were used as evidence-synthesis categories to organize the literature and generate candidate indicators, not as the final validated RAMP-ACE structure. Table 3 summarizes these preliminary domains, representative indicator statements, auditable evidence, reliability cues, and their mapping to the final RAMP-ACE framework. Most domains mapped directly to final framework areas, while standardization/calibration and quality assurance/continuous improvement were later consolidated into final D6, Standardization, calibration, and reflective QA.

Figure 5 shows that evidence coverage was strongest for assessment literacy and blueprinting, CBTA-aligned lesson design, and standardization/calibration, with more moderate support for laboratory safety pedagogy, human-performance learning, and facilitation/inclusion. Figure 6 indicates higher implementation readiness for lesson design, assessment literacy, standardization/calibration, and QA/continuous improvement, and moderate readiness for the remaining domains. Together, Table 3 and Figures 5–6 show that the SLR provided a structured preliminary evidence base for the Alignment Engine, candidate indicator pool, and pre-Delphi framework drafting. Subsequent regulatory triangulation and Delphi refinement consolidated overlapping constructs and produced the final six-domain RAMP-ACE framework.

Figure 5

Cross-study coverage of preliminary SLR-derived AMT domains and evidence types

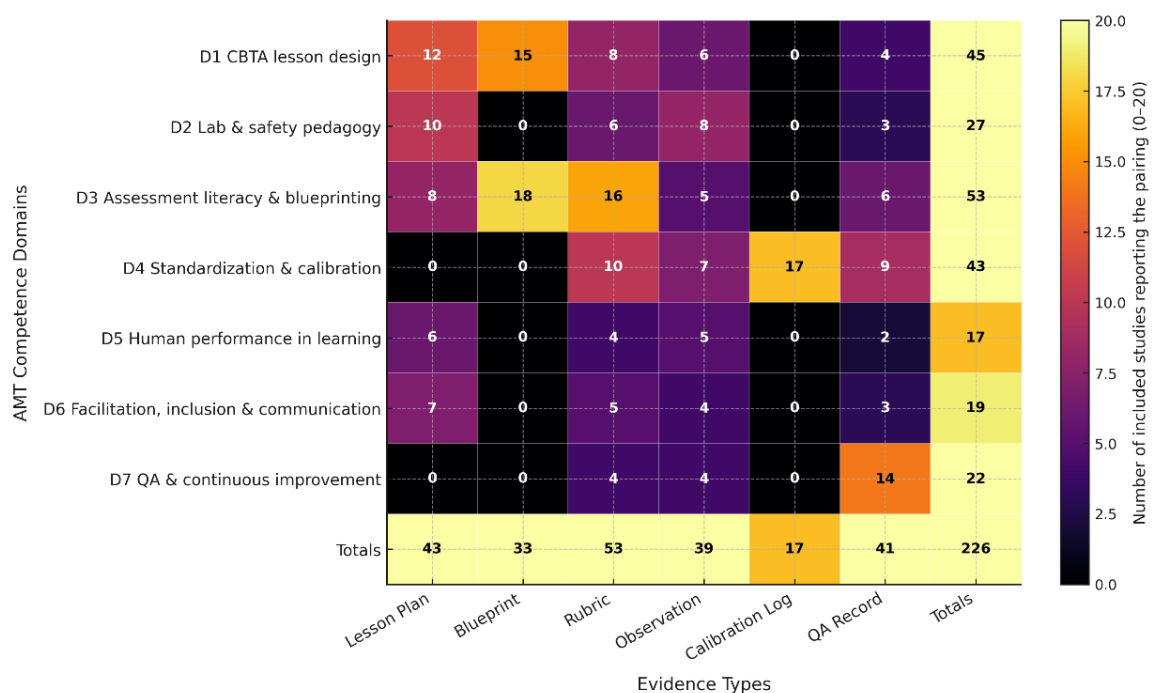
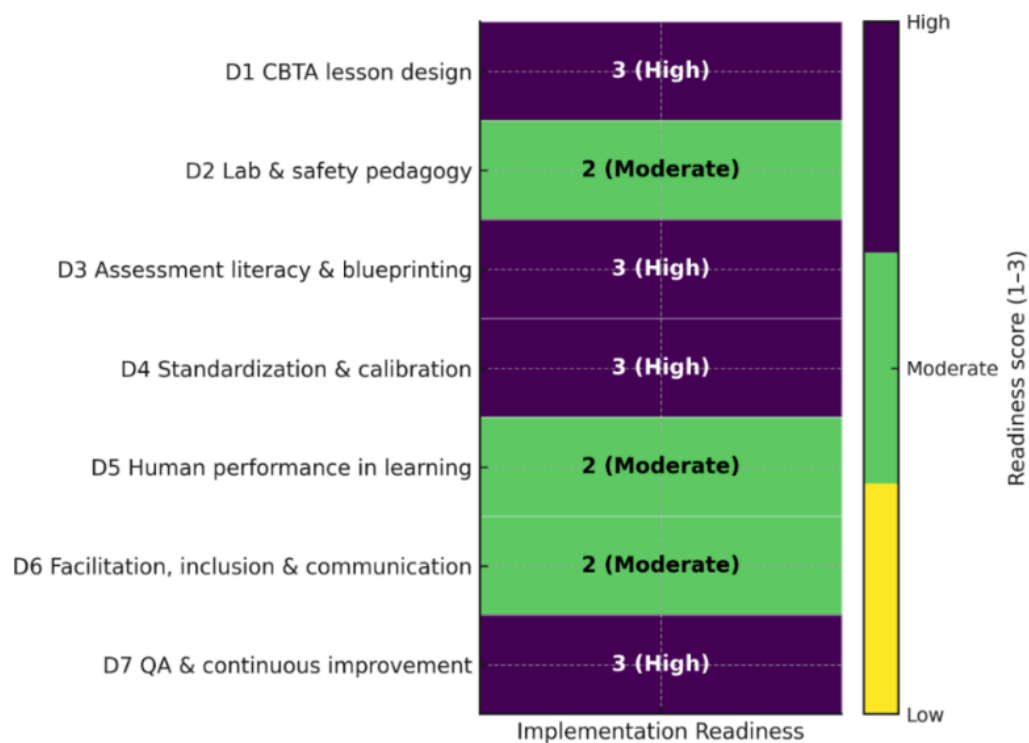


Table 3

Preliminary SLR-derived domains, indicator statements, auditable evidence, and final RAMP-ACE mapping (AMT)

Preliminary SLR-derived domain	Final RAMP-ACE mapping	Indicator (short code)	Indicator statement (observable)	Typical auditable evidence	Reliability cue
SLR-D1 CBTA lesson design	Final D1 CBTA-aligned lesson design	D1.1-Align	Lesson outcomes map 1:1 to tasks and standards (PCAR/CBTA).	Outcome–task mapping sheet; syllabus module map.	Alignment sheet signed-off.
SLR-D1 CBTA-aligned lesson design	Final D1 CBTA-aligned lesson design	D1.2-Scaffold	Lesson follows demonstration → guided practice → independent performance with error-tolerant checkpoints.	Lesson plan with timed phases; instructor notes.	≥ 2 plans show the sequence.
SLR-D2 Laboratory and safety pedagogy	Final D2 Laboratory and safety pedagogy	D2.1-ToolCtl	Tool-control briefing is provided, with FOD prevention embedded in task steps.	Tool-control briefing script; FOD checklist.	Alignment sheet signed off.
SLR-D2 Laboratory and safety pedagogy	Final D2 Laboratory and safety pedagogy	D2.2-EM-Teach	Error-management teaching is used to help learners anticipate, contain, and learn from errors.	Pre-task hazard identification; post-task error log and debrief.	At least two plans show the sequence.
SLR-D3 Assessment literacy and blueprinting	Final D3 Assessment literacy and blueprinting	D3.1-Blueprint	Practical and written assessment items sample all intended outcomes at planned weights.	Assessment blueprint; item bank index.	Spot-check log for at least two sessions.
SLR-D3 Assessment literacy and blueprinting	Final D3 Assessment literacy and blueprinting	D3.2-RubricQual	Rubrics specify pass/fail anchors for safety-critical performance steps.	Rubric with critical-step anchors; exemplar work.	Anchor examples attached.
D4 Standardization & calibration	Final D6 Standardization, calibration, and reflective QA	D4.1-IRR	Assessor interrater reliability is documented on sample tasks.	Calibration log; IRR stats or agreement form.	Target κ /ICC ≥ 0.70.
SLR-D4 Standardization and calibration	Final D6 Standardization, calibration, and reflective QA	D4.2-Norming	Pre-exam norming uses annotated exemplars and consensus scoring rules.	Norming pack; consensus minutes.	Sign-in sheet and version control.
SLR-D5 Human performance in learning	Final D4 Human-performance learning	D5.1-HP-Skills	Briefing teaches human-performance skills, including workload, fatigue, communication, and CRM.	Human-performance briefing slide; learner reflection.	Reflection samples attached.
SLR-D5 Human performance in learning	Final D4 Human-performance learning	D5.2-SafetyCase	Structured post-task debrief prompts learners to explain what happened, why it happened, and what should change next.	Debrief form; corrective actions log.	Close-out within 7 days.
SLR-D6 Facilitation, inclusion, and instructional communication	Final D5 Facilitation, inclusion, and instructional communication	D6.1-Access	Directions and safety information are accessible through multimodal and plain-language formats.	Quick-ref cards; signage photos.	Readability/UDL check noted.
SLR-D6 Facilitation, inclusion, and instructional communication	Final D5 Facilitation, inclusion, and instructional communication	D6.2-Feedback	Feedback is timely, specific, and linked to rubric criteria.	Feedback snapshots; timestamped LMS logs.	Median turnaround ≤ 72h.
SLR-D7 Quality assurance and continuous improvement	Final D6 Standardization, calibration, and reflective QA	D7.1-QA-Audit	Internal QA audits trace from outcome → rubric → evidence.	QA audit checklist; CARs.	Nonconformities tracked.
SLR-D7 Quality assurance and continuous improvement	Final D6 Standardization, calibration, and reflective QA	D7.2-CPD	Instructors log CBTA-related CPD with documented practice change.	CPD log + change memo.	At least two CPD activities linked to one practice change per year.

Figure 6*Implementation readiness of preliminary SLR-derived AMT domains*

3. Candidate framework v1 → v2 before final consolidation: domain definitions, indicators, proficiency anchors, evidence ledger, and regulatory mapping

The candidate framework was refined from v1 to v2 before final consolidation. Framework v1 contained 14 candidate indicators organized under the seven preliminary SLR-derived domains, while Framework v2 expanded the candidate pool to 23 indicators to strengthen coverage in assessment blueprinting, assessor calibration, laboratory and safety pedagogy, and human-performance learning. These seven domains were used as evidence-synthesis and candidate-development categories only, not as the final validated RAMP-ACE structure. Figure 7 summarizes the distribution of candidate indicators across the preliminary SLR-derived domains, while Figure 8 presents their pre-final regulatory alignment with PCAR and selected international references.

To support consistent interpretation, v2 specified proficiency anchors for each candidate domain using three performance levels: Novice, Competent, and Proficient/Calibrated. As shown in Table 4, these anchors translated each domain into observable instructional behaviors linked to auditable artifacts, including lesson plans, outcome–task maps, assessment blueprints, rubrics, calibration/IRR logs, structured debrief forms, and quality-assurance records. The evidence ledger further supported

traceability by linking each indicator to specific artifacts, artifact currency, reviewer notes, and regulatory or international reference points.

Regulatory mapping showed the strongest alignment for lesson design, assessment literacy, and standardization/calibration; broader support for laboratory safety pedagogy and human-performance learning; and more variable support for facilitation, inclusion, and quality-assurance constructs. These mapping results informed, but did not constitute, the final validated framework. During regulatory triangulation and Delphi refinement, overlap between standardization/calibration and quality assurance/continuous improvement was resolved by consolidating these areas into a single final domain. Two indicators were also removed for failing Delphi retention thresholds. Accordingly, the final validated RAMP-ACE framework consists of six domains and 21 retained indicators.

Figure 7

Candidate indicators per preliminary SLR-derived domain: framework v1 versus v2 before final consolidation

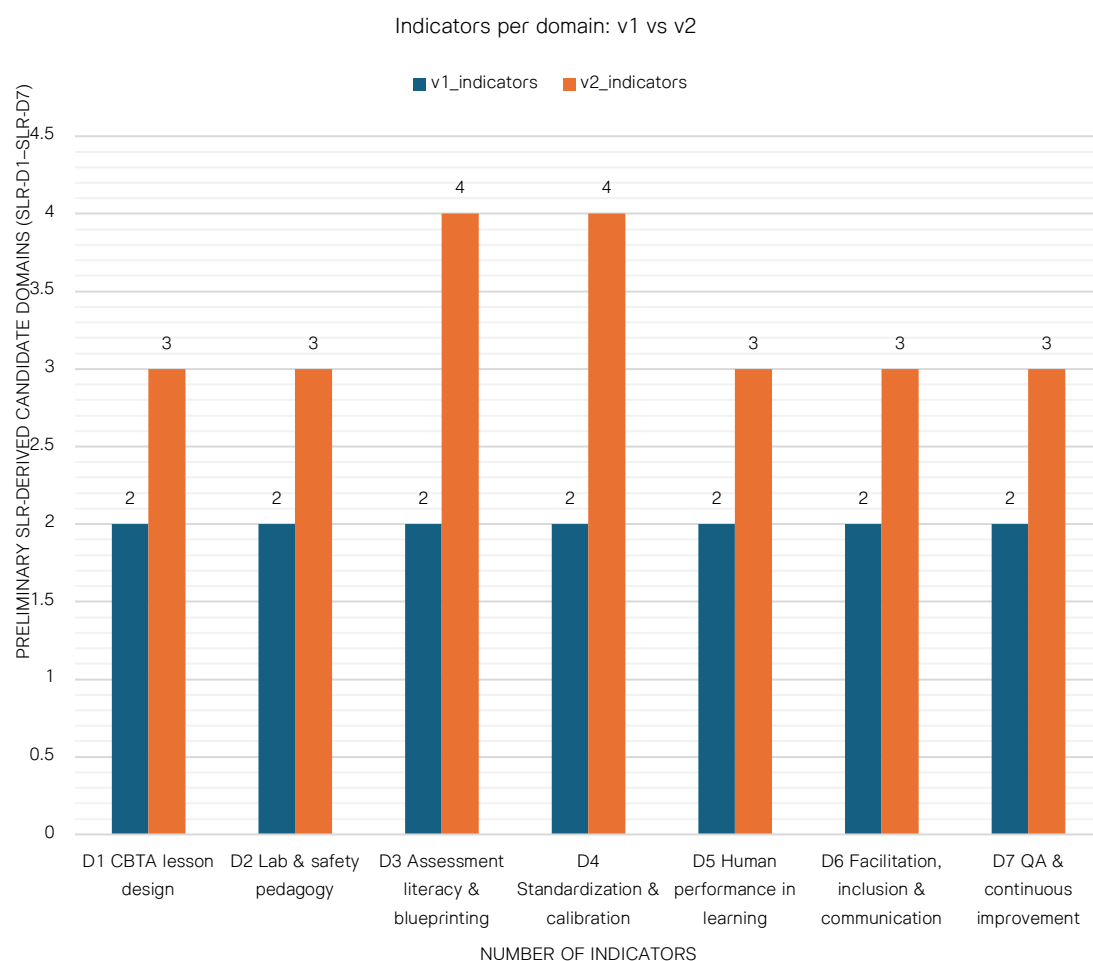
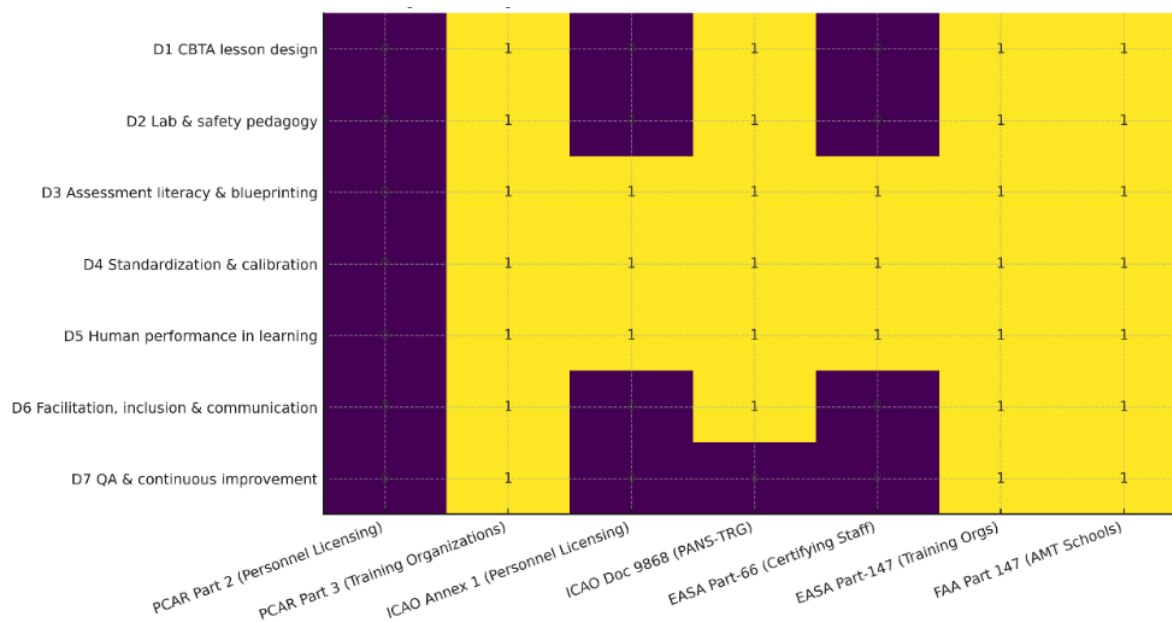


Figure 8*Pre-final alignment of candidate AMT domains with PCAR and international references***Table 4***Final RAMP-ACE proficiency anchors by validated domain*

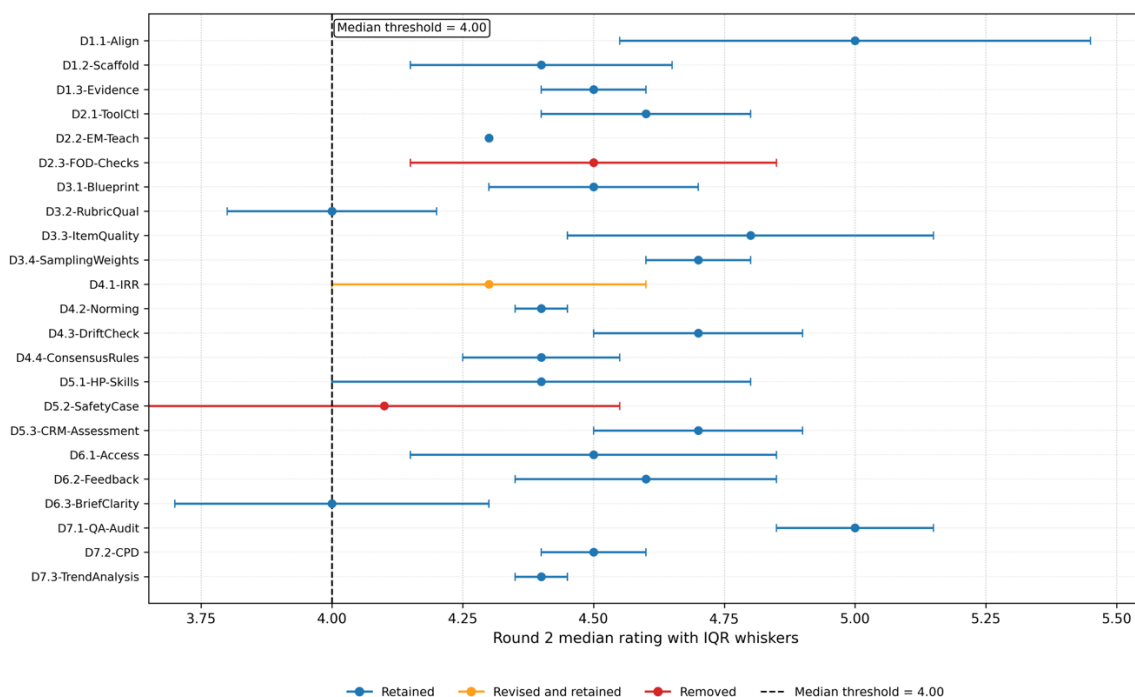
Final RAMP-ACE domain	Novice	Competent	Proficient/Calibrated
D1 CBTA-aligned lesson design	Outcomes are listed but weakly mapped to tasks; lesson sequencing is mostly ad hoc.	Outcomes are mapped to tasks and standards; lessons follow demonstration, guided practice, and independent performance.	Outcome-task mapping is validated; sequencing is revised using QA feedback; exemplars are archived.
D2 Laboratory and safety pedagogy	Safety briefings and tool-control practices are inconsistent or informal.	Standard safety briefings are used; tool control, hazard identification, and error logs are embedded in lab tasks.	Structured risk debriefs are used; recurring safety or error trends are tracked and closed through the QA cycle.
D3 Assessment literacy and blueprinting	Rubrics are generic; assessment coverage is incomplete or weakly documented.	Assessment blueprints cover intended outcomes; rubrics define safety-critical performance anchors.	Blueprints are audited; item and rubric quality evidence is filed; sampling weights are traceable.
D4 Human-performance learning	Human-performance topics are incidental, with limited learner reflection.	Human-performance briefings are integrated into instruction; guided reflections are collected.	Human-performance skills are assessed; reflection and debrief outputs are tracked to improvement actions.
D5 Facilitation, inclusion, and instructional communication	Instructions are dense; feedback is delayed, generic, or weakly linked to criteria.	Materials use plain-language and multimodal formats; feedback is timely and linked to rubric criteria.	Accessibility checks are documented; feedback analytics meet service targets and inform instructional improvement.
D6 Standardization, calibration, and reflective QA	Assessors work independently; IRR evidence, audit trails, and improvement records are limited or scattered.	Norming uses exemplars; agreement is documented; QA checks link outcomes, rubrics, evidence, and corrective actions.	IRR ≥ 0.70 is sustained; calibration packs are versioned; drift checks, QA trends, CPD, and corrective actions drive measurable practice change.

4. Delphi outcomes: medians/IQRs, content validity, and indicator decisions

The Delphi panel comprised 14 experts in Round 1 and 13 in Round 2, with one attrition between rounds. Consensus was stable, as median shifts across items were ≤ 0.5 . In Round 2, all 23 v2 candidate indicators achieved median ratings of at least 4.00, with IQRs ≤ 1.00 , indicating acceptable agreement on clarity, relevance, and necessity. Figure 9 shows the Round 2 medians and IQRs, with the median retention threshold marked at 4.00.

Figure 9

Round 2 Delphi median ratings and IQRs for v2 candidate indicators



Item-level content validity further differentiated retained from non-retained indicators. Most indicators met the I-CVI threshold of ≥ 0.78 , and the overall scale demonstrated strong content validity, with S-CVI/Ave = 0.89 and S-CVI/UA = 0.35, as summarized in Figure 10. Two candidate indicators, D2.3-FOD-Checks and D5.2-SafetyCase, achieved acceptable median and IQR values but fell below the I-CVI retention threshold and were therefore removed. D4.1-IRR was retained after revision because its median and IQR met consensus criteria, and the revision clarified calibration procedures and evidence expectations.

Table 5 reports the complete Delphi decision trail, including candidate codes, final RAMP-ACE codes, Round 2 medians, IQRs, I-CVI values, and final decisions. This mapping preserves traceability from the pre-final seven-domain candidate structure

to the validated six-domain RAMP-ACE framework. Overall, 20 indicators were retained without major revision, one indicator was revised and retained, and two indicators were removed, as summarized in Figure 11. The final validated framework therefore retained 21 indicators across six domains.

Figure 10

Item-level content validity index (I-CVI) for v2 candidate indicators

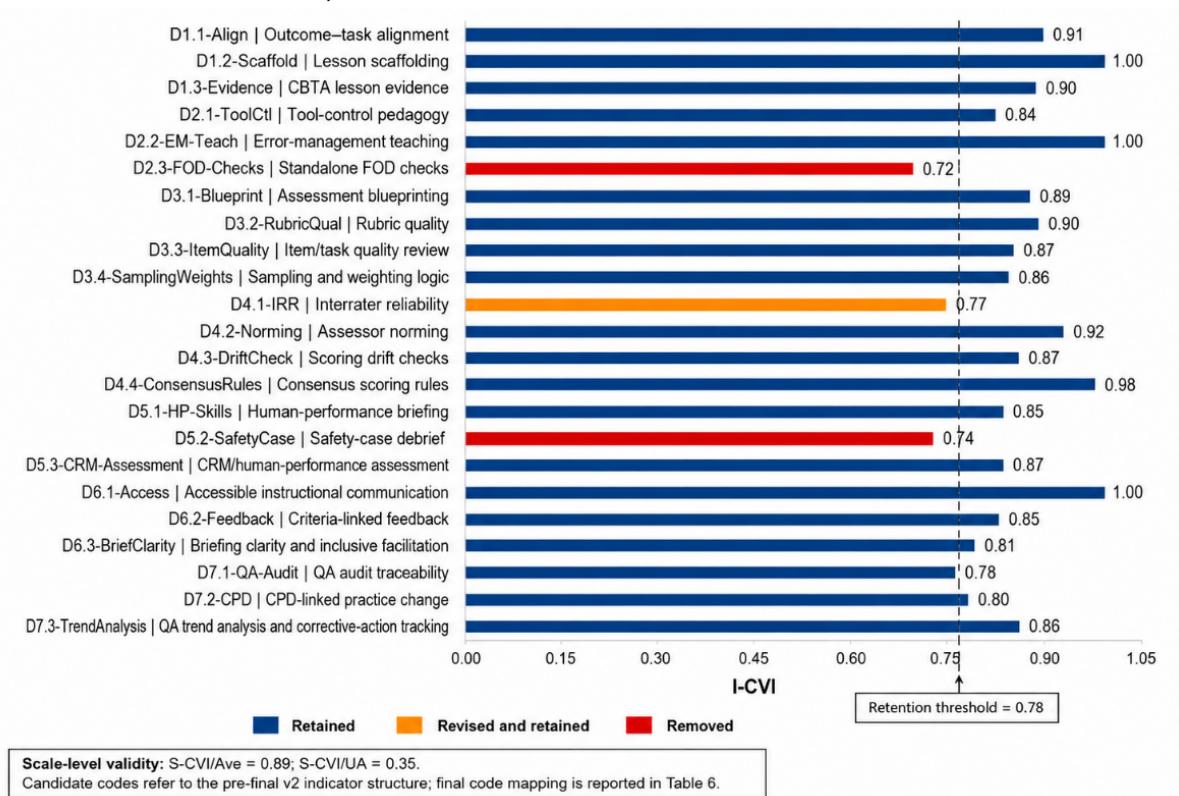


Figure 11

Summary of Delphi indicator decisions (Retained/Revised/Removed).

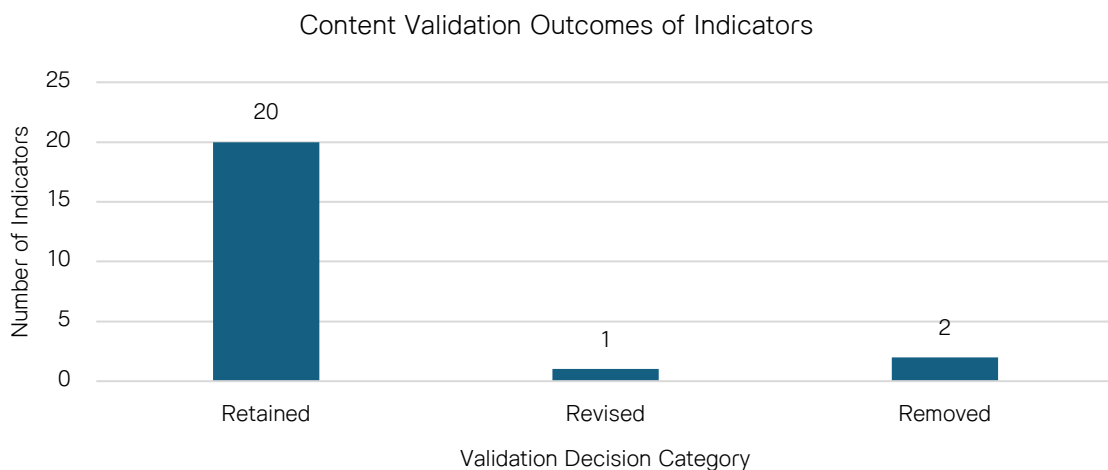


Table 5

Delphi statistics, decisions, and candidate-to-final RAMP-ACE code mapping.

Item Code	Final code	Median_R2	IQR_R2	I-CVI	Decision
D1.1-Align	Final D1.1	5.00	0.90	0.91	Retained
D1.2-Scaffold	Final D1.2	4.40	0.50	1.00	Retained
D1.3-Evidence	Final D1.3	4.50	0.20	0.90	Retained
D2.1-ToolCtl	Final D2.1	4.60	0.40	0.84	Retained
D2.2-EM-Teach	Final D2.2	4.30	0.00	1.00	Retained
D2.3-FOD-Checks	-	4.50	0.70	0.72	Removed
D3.1-Blueprint	Final D3.1	4.50	0.40	0.89	Retained
D3.2-RubricQual	Final D3.2	4.00	0.40	0.90	Retained
D3.3-ItemQuality	Final D3.3	4.80	0.70	0.87	Revised & Retained
D3.4-SamplingWeights	Final D3.4	4.70	0.20	0.86	Retained
D4.1-IRR	Final D6.1	4.30	0.60	0.77	Revised
D4.2-Norming	Final D6.2	4.40	0.10	0.92	Retained
D4.3-DriftCheck	Final D6.3	4.70	0.40	0.87	Retained
D4.4-ConsensusRules	Final D6.4	4.40	0.30	0.98	Retained
D5.1-HP-Skills	Final D4.1	4.40	0.80	0.85	Retained
D5.2-SafetyCase	-	4.10	0.90	0.74	Removed
D5.3-CRM-Assessment	Final D4.2	4.70	0.40	0.87	Retained
D6.1-Access	Final D5.1	4.50	0.70	1.00	Retained
D6.2-Feedback	Final D5.2	4.60	0.50	0.85	Retained
D6.3-BriefClarity	Final D5.3	4.00	0.60	0.81	Retained
D7.1-QA-Audit	Final D6.5	5.00	0.30	0.78	Retained
D7.2-CPD	Final D6.6	4.50	0.20	0.80	Retained
D7.3-TrendAnalysis	Final D6.7	4.40	0.10	0.86	Retained

5. ACE Cycle feasibility

The pilot ACE cycle demonstrated strong feasibility across multiple dimensions. Interrater agreement (IRR) on rubric-based dossier scoring met the pre-specified target of $\kappa \geq 0.70$, with coefficients ranging from 0.72 to 0.86, confirming calibration consistency among assessors (Figure 12). The time-to-rate per dossier was manageable within instructional schedules, with a median of 38 minutes (IQR 12) across raters (Figure 13). Artifact availability was high for blueprinting, rubrics, and calibration evidence (> 85%), but substantially lower for structured debriefs (62%) and human-performance reflections (58%), indicating specific documentation gaps that require attention (Figure 14).

The corrective actions log further underscored areas for refinement. Torque application steps were reinforced with torque-card prompts; fault-isolation sequencing

was clarified with decision-tree cues; rivet spacing was safeguarded with go/no-go gauges; and leak testing was standardized with stabilization timers. Each action was assigned to a responsible owner with a closure deadline, and most were resolved within 2 weeks (Table 6).

Together, these findings confirm that the ACE cycle is operationally viable and scalable, while also identifying domains—particularly structured debriefs and human-performance documentation—that require targeted strengthening prior to broader implementation. Presenting interrater reliability, dossier scoring efficiency, artifact availability, and corrective actions in a unified results panel (Figures 12–14, Table 6) ensures transparent reporting of feasibility outcomes.

Figure 12

Interrater agreement by task (target ≥ 0.70)

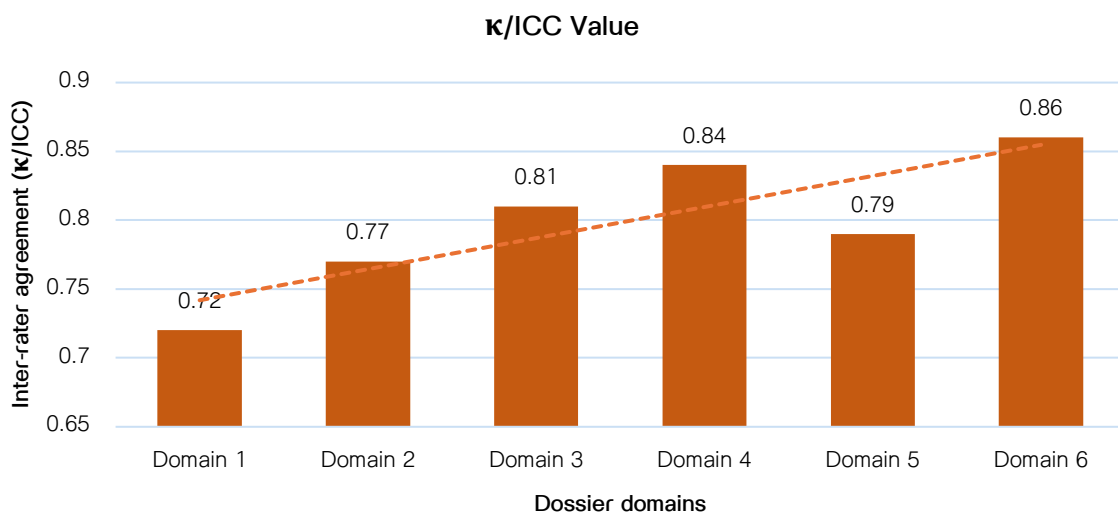


Figure 13

Median time-to-rate per dossier across assessors (whiskers = $\pm 0.5 \cdot IQR$).

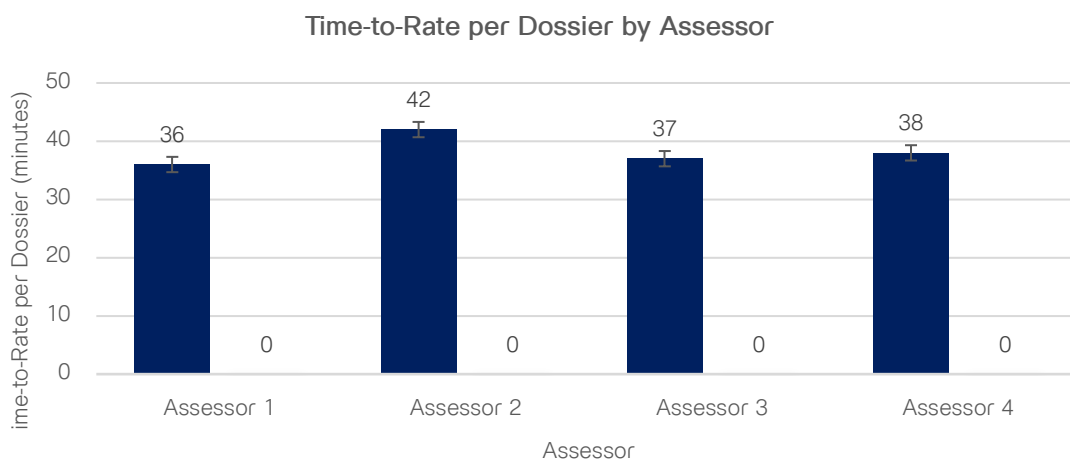


Figure 14

Artifact availability by indicator during ACE pilot.

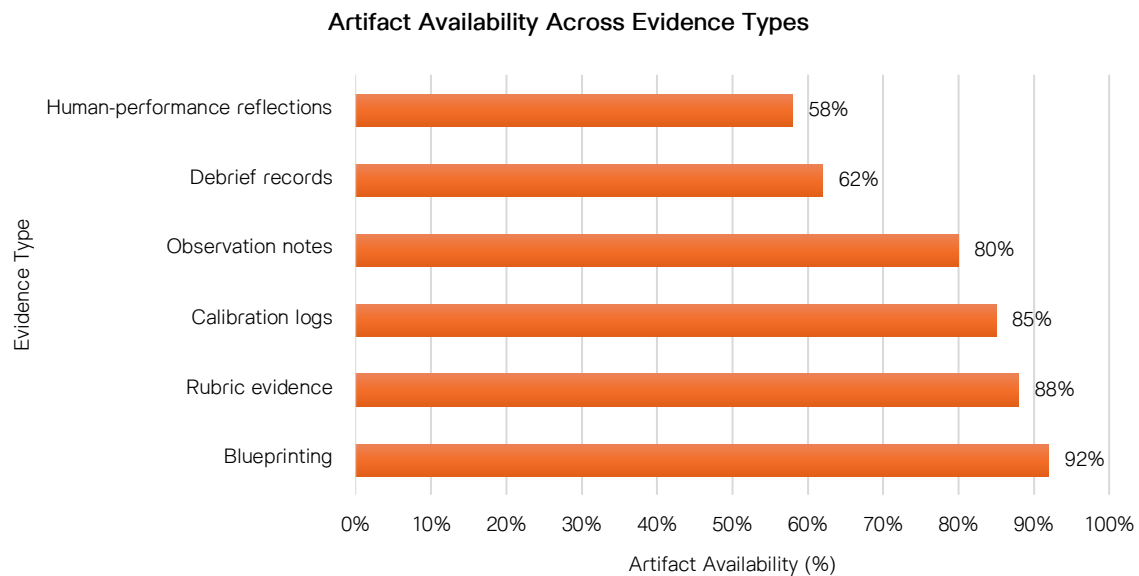


Table 6

Corrective actions identified in the ACE pilot cycle, with owners and closure timelines.

Issue observed	Corrective action	Owner	Status
Torque application missteps	Add torque-card step; reinforce click-type wrench reading	Lead Instructor	Closed (7 days)
Fault isolation skipped	Insert decision-tree prompt; add continuity demo	Module Owner	Open (14 days)
Rivet edge distance below minimum	Introduce go/no-go gauge; use pre-drill templates	QA Rep	Closed (3 days)
Premature manometer reading	Add stabilization timer and acceptance criteria prompt	ATC Liaison	Closed (10 days)

Discussion

1. What RAMP-ACE Changes

RAMP-ACE replaces the long-standing assumption that technical licensing equates to teaching competence with a framework that is explicitly structured around six RAMP-ACE domains, each defined by observable behaviors, calibratable indicators, and auditable evidence. Instructional competence is operationalized through six domains, associated behavioral indicators, proficiency anchors, and required artifacts, making pedagogical performance verifiable through lesson plans, assessment blueprints, rubrics, observation notes, calibration logs, and quality-assurance records.

The framework drives improvement through three mechanisms. First, it integrates constructive-alignment principles by linking learning outcomes, instructional strategies, and assessments across the six-domain structure, reducing hidden curriculum effects and clarifying performance expectations. Second, it standardizes evaluation practices through shared scoring anchors and IRR/ICC targets applied consistently across all six domains, decreasing rater variation and supporting cross-instructor comparability. Third, it embeds evidence-based instructional quality assurance by mapping each domain indicator to specific artifacts, enabling objective oversight and traceability within Aviation Training Organizations (ATOs).

Findings demonstrate strong overall validity and feasibility, including high scale-level content validity indices (S-CVI), κ /ICC values ≥ 0.70 , acceptable rating times, and high artifact availability. However, results were mixed in selected domains where regulatory guidance was strong (e.g., CBTA-aligned lesson design and assessment validity). Yet, Delphi agreement was comparatively lower, indicating areas where regulatory expectations remain conceptually ambiguous for instructors. These domain-specific discrepancies informed targeted refinements to indicator wording, anchoring precision, and definitions of evidence during framework consolidation.

For regulators and ATOs, RAMP-ACE translates broad policy mandates into operational controls—such as standardized blueprints, calibration packages, and documented corrective-action closures—supporting consistent, data-driven remediation across the six-domain framework. For learners, instruction becomes risk-focused and system-reliable, shifting from personality-driven teaching to transparent, evidence-linked pedagogy. Remaining priorities include strengthening documentation practices, expanding calibration cycles, and enabling automated evidence capture. Collectively, RAMP-ACE reframes the instructor role from “licensed expert” to accountable educator operating within a clearly bounded six-domain competence model anchored in demonstrable instructional performance rather than assumption.

2. Policy Implications for PCAR Part 3

Validation evidence—including high scale-level content validity (S-CVI/Ave), interrater reliability coefficients (κ /ICC ≥ 0.70), and strong artifact availability—supports targeted amendments to PCAR Part 3 Instructor Requirements. Specifically, the following provisions are recommended: (g) outcome-to-task alignment supported by module-level assessment blueprinting; (h) maintained scoring rubrics with clearly defined critical-step anchors; (i) documented participation in calibration activities demonstrating κ /ICC ≥ 0.70 or an approved remediation plan; (j) a module-level instructional evidence ledger (lesson plans, assessment blueprints, rubrics, observation notes, calibration logs, quality-assurance records, structured debriefs,

and human-performance reflections); and (k) completion of at least one documented Assessment–Calibration–Enhancement (ACE) cycle per academic year.

These additions operationalize the six consolidated RAMP-ACE domains by converting the implicit assumption that “licensed expert = effective instructor” into an auditable, evidence-based instructional standard. They reinforce domains with consistently high consensus (e.g., CBTA-aligned lesson design, assessment literacy, and standardization with reflective practice) while explicitly addressing recurrent gaps identified in the SLR and Delphi phases, particularly in structured debriefing and human-performance-oriented reflection.

To ensure regulatory clarity and uniform implementation, an Implementing Standard or Annex to PCAR Part 3 should formally codify the six RAMP-ACE domains, required evidence types, the ACE quality-assurance cycle, and maturity thresholds. A four-level progression is recommended: L1 (Baseline)—required artifacts complete; L2 (Standardized)—quality-assured blueprints and rubrics in routine use; L3 (Calibrated)—sustained $\kappa/\text{ICC} \geq 0.70$ with periodic drift checks; and L4 (Assured)—closed-loop quality assurance with documented instructional improvements. Standardized templates (e.g., blueprint matrices, calibration logs, debrief forms) would further streamline ATO reporting and reduce compliance burden.

A phased regulatory rollout over 12–24 months is advised: publication of templates and guidance (0–6 months); shadow ACE cycles and baseline data collection (6–12 months); and progressive compliance audits emphasizing calibration integrity and weak evidence classes (12–24 months). Equivalency recognition (e.g., ICAO TIC-aligned systems) may be granted where full coverage of the six RAMP-ACE domains and documented ACE completion are demonstrated. Recurrent standardization reviews should occur every 24 months, with targeted remediation triggered when κ/ICC values fall below 0.70.

3. ATO Implementation Guide

QA embedding. Integrate the ACE cycle into routine quality-assurance operations across the six RAMP-ACE domains. Conduct quarterly mini-audits covering at least 10% of modules, reviewing complete evidence ledgers (lesson plans, assessment blueprints, rubrics, observation notes, calibration logs, QA records, structured debriefs, and human-performance reflections). Schedule two peer observations per instructor each term, prioritizing lab and safety pedagogy and human-performance-related domains, where evidence maturity has been weakest. Semiannual standardization workshops should target $\text{IRR}/\text{ICC} \geq 0.70$, with drift checks and documented remediation as required. All corrective actions are recorded in a central register, with closure artifacts retained to ensure traceable improvement.

Data and targets. Use a streamlined QA dashboard to monitor: IRR/ICC trendlines (target ≥ 0.70 ; drift < 0.05 over 12 months); the proportion of instructors achieving maturity level $\geq L2$ across the six-domain framework (targets: $\geq 80\%$ at 12 months; $\geq 95\%$ at 24 months); and corrective-action closure time (median ≤ 14 days; 90th percentile ≤ 30 days). Maintain artifact availability $\geq 85\%$ for lesson design, assessment, and calibration evidence, and increase documentation of debriefing and human-performance reflection from approximately 60% to $\geq 80\%$ within the first year.

Cadence and ownership. QA leads oversee audits and dashboard monitoring; module owners manage peer observations and evidence collection; and a designated calibration coordinator facilitates workshops and tracks IRR/ICC performance. Quarterly ACE reviews address low agreement, delayed closures, or missing artifacts through targeted mini-cycles, ensuring consistent application of the six RAMP-ACE domains while sustaining scoring reliability, efficient evaluation, and progressive documentation maturity.

4. Global and TVET Transferability

RAMP-ACE functions as a modular, evidence-based validation architecture that can sit on top of existing training rules rather than replacing them. Its validation profile—high S-CVI/Ave, acceptable S-CVI/UA, κ /ICC ≥ 0.70 , and strong artifact availability—shows that a combined PRISMA–Delphi–IRR sequence can deliver a robust, multi-evidence pipeline for developing instructional quality-assurance frameworks. Although instantiated in aviation, this methodology is transferable to TVET and other regulated training systems that require evidence-based instructional quality assurance, standardized assessment practices, and reproducible calibration processes. The core building blocks—domains, indicators, evidence ledgers, and the ACE cycle—parallel QA mechanisms already used in fields such as maritime, mechatronics, engineering technologies, and allied health.

Cross-sector adoption can follow a staged pathway: introduce the six RAMP-ACE domains with a minimum evidence ledger; implement annual ACE cycles and maturity thresholds (L1–L4); and require recurrent standardization with drift checks and targeted mini-cycles whenever κ /ICC falls below the agreed threshold. Because RAMP-ACE is artifact-based, alignment with different jurisdictions or sectors becomes a documentation exercise rather than a full regulatory rewrite; a simple mapping template can link each indicator to local standards and required evidence. This makes the framework suitable not only as a regulatory tool but also as a methodological model for designing, validating, and maintaining teaching frameworks, rubrics, and accreditation instruments across diverse TVET and professional education contexts.

Limitations and Future Directions

This study has several limitations. First, RAMP-ACE was validated in aviation maintenance training using a bounded regulatory, documentary, and expert-consensus design; therefore, its generalizability to other technical education settings requires further testing. Second, the Delphi panel provided strong content-validity evidence, but the results remain dependent on expert judgment and may reflect the composition of the selected panel. Third, the pilot audit assessed feasibility, artifact availability, rating time, and interrater reliability, but did not directly measure long-term effects on instructor practice, learner competence, or safety outcomes. Finally, the use of instructor dossiers strengthened auditability but may underrepresent informal instructional behaviors that are not routinely documented.

Future research should conduct cross-institutional validation of the final six-domain RAMP-ACE framework across diverse ATOs and technical education providers. Further studies should test the stability of the D1–D6 rubric across raters, institutions, course types, and regulatory contexts. Longitudinal research is also needed to examine whether improved scores in D1 CBTA-aligned lesson design, D2 laboratory and safety pedagogy, D3 assessment literacy and blueprinting, D4 human-performance learning, D5 facilitation, inclusion, and instructional communication, and D6 standardization, calibration, and reflective QA are associated with stronger instructional quality, learner competence, audit readiness, and corrective-action closure. Future work may also develop a digital evidence ledger and dashboard to support routine monitoring, calibration, and continuous improvement within ATO quality systems.

Conclusion

PCAR Part 3 effectively governs the technical what of Aircraft Maintenance Technology (AMT) training but offers limited guidance on the pedagogical how, contributing to weak audit mechanisms and variable instructional quality. This study addressed that gap by developing and validating RAMP-ACE, a regulatory-aligned, evidence-based framework structured around six consolidated pedagogical domains, with defined indicators, proficiency anchors, and an auditable evidence ledger. Validation results—including high S-CVI/Ave values, $\kappa/\text{ICC} \geq 0.70$, manageable rating times, and strong artifact availability—confirm both content validity and operational feasibility, while also identifying targeted gaps in structured debriefing and human-performance–focused reflection.

A central contribution of this work is its integrated, multi-source validation architecture, which extends beyond Delphi-only or perception-based approaches common in educational research. By combining regulatory document analysis, systematic literature synthesis, structured expert consensus, and empirical reliability testing, the study demonstrates a more rigorous pathway for validating instructional competence. This architecture operationalizes quality assurance, calibration, and evidence traceability within a single framework, responding to long-standing calls for auditable, system-level validation in TVET and other safety-critical training environments. Although demonstrated in aviation, the approach is readily transferable to broader regulated education contexts that require standardized, defensible competency systems.

Adoption within aviation can proceed through non-rule mechanisms, such as an Implementing Standard or Annex codifying the six RAMP-ACE domains, associated evidence requirements, the Assessment–Calibration–Enhancement (ACE) cycle, and maturity thresholds (L1–L4), supported by recurrent standardization and drift monitoring. Together, these measures shift instructional oversight from assumption-based credentialing toward demonstrable, evidence-linked teaching competence, strengthening regulatory assurance, supporting ATO quality systems, and improving the consistency, fairness, and safety value of AMT instruction.

Author Contributions The sole author was responsible for the study conception and design, data collection, analysis, interpretation, manuscript drafting, revision, and final approval of the submitted version.

Competing Interests The author declares that there are no competing interests related to this study.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability The data supporting this study consist of systematic-review screening records, extracted evidence tables, Delphi summary ratings, and anonymized pilot-audit summaries. These materials are available from the author upon reasonable request, subject to confidentiality and ethical restrictions. No publicly archived dataset was generated.

Declaration of AI and Tool Use

Grammarly, Microsoft Editor, and ChatGPT (OpenAI) were used only for grammar, clarity, phrasing, and readability improvements. Microsoft Excel and Canva were used to prepare and format figures. No AI tool was used to generate data, conduct analysis, or create interpretations. All content, analyses, results, and visuals were reviewed and verified by the author.

References

- Abildgren, L., Lebahn-Hadidi, M., Mogensen, C. B., Toft, P., Nielsen, A. B., Frandsen, T. F., Steffensen, S. V., & Hounsgaard, L. (2022). The effectiveness of improving healthcare teams' human factor skills using simulation-based training: A systematic review. *Advances in Simulation*, 7(1), Article 12. <https://doi.org/10.1186/s41077-022-00207-2>
- Basileo, L. D., & Toth, M. (2019). A state level analysis of the Marzano Teacher Evaluation Model: Predicting teacher value-added measures with observation scores. *Practical Assessment, Research & Evaluation*, 24(1), Article 6. <https://doi.org/10.7275/cc5b-6j43>
- Chen, A. M. H., Kleppinger, E. L., Churchwell, M. D., & Rhoney, D. H. (2024). Examining competency-based education through the lens of implementation science: A scoping review. *American Journal of Pharmaceutical Education*, 88(2), Article 100633. <https://doi.org/10.1016/j.ajpe.2023.100633>
- Chen, S., Zhang, C., & Li, W. (2022). The effects of competency-based training model in the training of new nurses: A meta-analysis and systematic review. *PLOS ONE*, 17(11), Article e0277484. <https://doi.org/10.1371/journal.pone.0277484>
- Clare, J., & Kourousis, K. I. (2021). Learning from incidents in aircraft maintenance and continuing airworthiness: Regulation, practice and gaps. *Aircraft Engineering and Aerospace Technology*, 93(2), 338–346. <https://doi.org/10.1108/aeat-06-2020-0114>
- Edgar, A., & Stack, K. M. (2023). Parallel incorporation and public law. *International Journal of Constitutional Law*, 21(3), 734–769. <https://doi.org/10.1093/icon/moad064>
- Edge, C., Monske, E., Boyer-Davis, S., VandenAvond, S., & Hamel, B. (2021). Leading university change: A case study of meaning-making and implementing online learning quality standards. *American Journal of Distance Education*, 36(1), 53–69. <https://doi.org/10.1080/08923647.2021.2005414>
- European Union Aviation Safety Agency. (2023). *AMC & GM to Part-147—Issue 2, Amendment 3*. https://www.easa.europa.eu/sites/default/files/dfu/amc_gm_to_part-147_-_issue_2_amendment_3.pdf
- Federal Aviation Administration. (2022). *Aviation instructor's handbook (FAA-H-8083-9B)*. U.S. Department of Transportation. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/aviation_instructors_handbook
- Franciosi, C., Di Pasquale, V., Iannone, R., & Miranda, S. (2019). A taxonomy of performance shaping factors for human reliability analysis in industrial maintenance. *Journal of Industrial Engineering and Management*, 12(1), 115–132. <https://doi.org/10.3926/jiem.2702>
- Haapamäki, E. (2022). Insights into neo-institutional theory in accounting and auditing regulation research. *Managerial Auditing Journal*, 37(3), 336–357. <https://doi.org/10.1108/maj-10-2020-2864>

- Heinonen, H., Virtaluoto, J., Suomivuori, T., Forsman, K., Kangas, T., & Siltanen, S. (2022). Minimalism for the win: User-centered design for guidance in industrial maintenance. *IEEE Transactions on Professional Communication*, 65(4), 485–501. <https://doi.org/10.1109/tpc.2022.3205468>
- Hovlid, E., Braut, G. S., Hannisdal, E., Walshe, K., Bukve, O., Flottorp, S., Stensland, P., & Frich, J. C. (2020). Mediators of change in healthcare organizations subject to external assessment: A systematic review with narrative synthesis. *BMJ Open*, 10(8), Article e038850. <https://doi.org/10.1136/bmjopen-2020-038850>
- International Air Transport Association, & International Federation of Air Line Pilots' Associations. (2021). *Guidance material and best practices for instructor and evaluator training* (2nd ed.). <https://www.ifalpa.org/publications/guidance-material-for-instructor-and-evaluator-training-edition-2-2633/>
- International Air Transport Association. (2025). *Competency assessment and evaluation for pilots, instructors and evaluators: Guidance material*. <https://www.iata.org/contentassets/c0f61fc821dc4f62bb6441d7abedb076/competency-assessment-and-evaluation-for-pilots-instructors-and-evaluators-gm.pdf>
- International Civil Aviation Organization. (2020). *Procedures for air navigation services: Training* (Doc 9868; 3rd ed.). <https://store.icao.int/en/procedures-for-air-navigation-services-training-doc-9868>
- International Civil Aviation Organization. (2025). *Training instructors course (TIC)*. <https://store.icao.int/en/training-instructors-course-tic-en>
- James, H. K., Chapman, A. W., Pattison, G. T. R., Fisher, J. D., & Griffin, D. R. (2020). Analysis of tools used in assessing technical skills and operative competence in trauma and orthopedic surgical training. *JBJS Reviews*, 8(6), Article e19.00167. <https://doi.org/10.2106/jbjs.rvw.19.00167>
- Jiménez Sierra, Á. A., Ortega Iglesias, J. M., Cabero-Almenara, J., & Palacios-Rodríguez, A. (2023). Development of the teacher's technological pedagogical content knowledge (TPACK) from the lesson study: A systematic review. *Frontiers in Education*, 8, Article 1078913. <https://doi.org/10.3389/feduc.2023.1078913>
- Karunakaran, C. S., & Ashok Babu, J. (2022). Impact of human factors in aircraft accident mitigation and aircraft maintenance training needs in post COVID-19 aviation. *Aircraft Engineering and Aerospace Technology*, 94(8), 1296–1302. <https://doi.org/10.1108/AEAT-10-2021-0300>
- Küçük Yılmaz, A. (2019). Strategic approach to managing human factors risk in aircraft maintenance organization: Risk mapping. *Aircraft Engineering and Aerospace Technology*, 91(4), 654–668. <https://doi.org/10.1108/AEAT-06-2018-0160>
- Mazzone, E., Puliatti, S., Amato, M., Bunting, B., Rocco, B., Montorsi, F., Mottrie, A., & Gallagher, A. G. (2020). A systematic review and meta-analysis on the impact of proficiency-based progression simulation training on performance outcomes. *Annals of Surgery*, 274(2), 281–289. <https://doi.org/10.1097/sla.0000000000004650>

- Mendes, N., Vieira, J. G. V., & Mano, A. P. (2022). Risk management in aviation maintenance: A systematic literature review. *Safety Science*, 153, Article 105810. <https://doi.org/10.1016/j.ssci.2022.105810>
- Mitchell, E. L. (2025). Assessment in the era of competency-based surgical education. *Seminars in Vascular Surgery*, 38(2), 122–131. <https://doi.org/10.1053/j.semvascsurg.2025.04.006>
- Morris-Mathews, H., Stark, K. R., Jones, N. D., Brownell, M. T., & Bell, C. A. (2021). Danielson's framework for teaching: Convergence and divergence with conceptions of effectiveness in special education. *Journal of Learning Disabilities*, 54(1), 66–78. <https://doi.org/10.1177/0022219420941804>
- Niederberger, M., Schifano, J., Deckert, S., Hirt, J., Homberg, A., Köberich, S., Kuhn, R., Rommel, A., & Sonnberger, M. (2024). Delphi studies in social and health sciences—Recommendations for an interdisciplinary standardized reporting (DELPHISTAR): Results of a Delphi study. *PLOS ONE*, 19(8), Article e0304651. <https://doi.org/10.1371/journal.pone.0304651>
- Paquibut, R., & Al Naamany, A. (2020). Managing organizational change to meet the research–teaching nexus standard: The case of an HEI in the Sultanate of Oman. *International Journal of Educational Management*, 34(4), 782–793. <https://doi.org/10.1108/IJEM-07-2019-0225>
- Peiró, J. M., Nielsen, K., Latorre, F., Shepherd, R., & Vignoli, M. (2020). Safety training for migrant workers in the construction industry: A systematic review and future research agenda. *Journal of Occupational Health Psychology*, 25(4), 275–295. <https://doi.org/10.1037/ocp0000178>
- Piggott, T., Langendam, M., Parmelli, E., Adolfsson, J., Akl, E. A., Armstrong, D., Braithwaite, J., Brignardello-Petersen, R., Brozek, J., Gore-Booth, J., Follmann, M., Leś, Z., Meerpohl, J. J., Neamțiu, L., Nothacker, M., Qaseem, A., Giorgi Rossi, P., Saz-Parkinson, Z., van der Wees, P., & Schünemann, H. J. (2021). Bringing two worlds closer together: A critical analysis of an integrated approach to guideline development and quality assurance schemes. *BMC Health Services Research*, 21, Article 172. <https://doi.org/10.1186/s12913-020-05819-w>
- Safari, K., McKenna, L., & Davis, J. (2023). The evolution of registration standards for internationally qualified midwives in Australia. *Midwifery*, 119, Article 103606. <https://doi.org/10.1016/j.midw.2023.103606>
- Sarfati, L., Ranchon, F., Vantard, N., Schwiertz, V., Larbre, V., Parat, S., Faudel, A., & Rioufol, C. (2019). Human-simulation-based learning to prevent medication error: A systematic review. *Journal of Evaluation in Clinical Practice*, 25(1), 11–20. <https://doi.org/10.1111/jep.12883>
- Scorgie, D., Feng, Z., Paes, D., Parisi, F., Yiu, T. W., & Lovreglio, R. (2024). Virtual reality for safety training: A systematic literature review and meta-analysis. *Safety Science*, 171, Article 106372. <https://doi.org/10.1016/j.ssci.2023.106372>

- Stievano, A., Caruso, R., Pittella, F., Shaffer, F. A., Rocco, G., & Fairman, J. (2019). Shaping nursing profession regulation through history: A systematic review. *International Nursing Review*, 66(1), 17–29. <https://doi.org/10.1111/incr.12449>
- Suharno, Pambudi, N. A., & Harjanto, B. (2020). Vocational education in Indonesia: History, development, opportunities, and challenges. *Children and Youth Services Review*, 115, Article 105092. <https://doi.org/10.1016/j.childyouth.2020.105092>
- Teresi, J. A., Yu, X., Stewart, A. L., & Hays, R. D. (2022). Guidelines for designing and evaluating feasibility pilot studies. *Medical Care*, 60(1), 95–103. <https://doi.org/10.1097/MLR.0000000000001664>
- Thulasy, T. N., Nohuddin, P. N. E., Nusyirwan, I. F., Rahim, N. A., Amrin, A., & Chua, S. (2022). Skills assessment criteria for aircraft maintenance technician in the context of Industrial Revolution 4.0. *Journal of Aerospace Technology and Management*, 14. <https://doi.org/10.1590/jatm.v14.1286>
- Tyagi, A., Tripathi, R., & Bouarfa, S. (2023). Learning from past in the aircraft maintenance industry: An empirical evaluation in the safety management framework. *Heliyon*, 9(11), Article e21620. <https://doi.org/10.1016/j.heliyon.2023.e21620>
- Vlasblom, J. I. D., Pennings, H. J. M., van der Pal, J., & Oprins, E. A. P. B. (2020). Competence retention in safety-critical professions: A systematic literature review. *Educational Research Review*, 30, Article 100330. <https://doi.org/10.1016/j.edurev.2020.100330>
- Vujičić, S., Hasanspahić, N., Gundić, A., & Hrdalo, N. (2020). Assessment for ensuring adequately qualified instructors in maritime education and training institutions. *Athens Journal of Sciences*, 7(2), 115–126. <https://doi.org/10.30958/ajs.7-2-4>
- Yusop, S. R. M., Rasul, M. S., Mohamad Yasin, R., Hashim, H. U., & Jalaludin, N. A. (2022). Assessment approaches and learning outcomes in technical and vocational education: A systematic review using PRISMA. *Sustainability*, 14(9), Article 5225. <https://doi.org/10.3390/su14095225>
- Yusop, S. R. M., Rasul, M. S., Mohammad Yasin, R., & Hashim, H. U. (2023). Identifying and validating vocational skills domains and indicators in classroom assessment practices in TVET. *Sustainability*, 15(6), Article 5195. <https://doi.org/10.3390/su15065195>