 ISSN 2579-7298

Development of a Web-Based Aviation English Proficiency Test: Integrating Adaptive Algorithms and Dynamic Assessment for Enhanced Evaluation in Aviation Education

Harunur Rosyid[1], Laila Rochmawati[2], Tiara Sylvia[3], Ahmad Rossydi[4], Henny Dwi Bhakti[5]

[1]*Universitas Muhammadiyah Gresik, Indonesia*
[2]*Politeknik Penerbangan Surabaya, Indonesia*
[3]*Politeknik Penerbangan Medan, Indonesia*
[4]*Politeknik Penerbangan Makassar, Indonesia*
[5]*Universitas Muhammadiyah Gresik, Indonesia*

ARTICLE INFO

Article history
Received
Revised
Accepted

Keywords
Aviation English
Proficiency Test
Adaptive Algorithms
Dynamic Assessment
Item Response Theory

ABSTRACT

Aviation English proficiency is pivotal for aviation school students to ensure secure communication in global airspace per ICAO guidelines. Conventional methods are rigid, leading to inaccurate and time-consuming evaluations that hinder training efficacy. This research develops a web-based adaptive Aviation English proficiency test integrating adaptive algorithms like Item Response Theory and dynamic assessment to enhance aviation education outcomes. Using a mixed-methods framework with the ADDIE model and quantitative experimental approach, an explanatory sequential design with non-equivalent control group was employed, involving needs assessment, prototype development, validation, and implementation. The sample included 141 aviation school students. Data from pre/post-tests were analyzed via SPSS. The findings showed that i) the web-based test is valid and feasible as an assessment tool with a validation score of 89.5%; ii) student proficiency levels are significantly improved before and after using the adaptive system (paired t-test: mean rise from 72.6 to 91.4, t=-14.28, p=0.000 <0.05); iii) dynamic assessment positively impacts learning outcomes following implementation (32% uplift, β=0.61, p<0.01); and iv) there is a significant difference between experimental and control groups in evaluation efficiency (independent t-test: 25% higher for experimental, t=10.52, p=0.000 <0.05). These affirm the test's efficacy, recommending broader adoption for refined aviation training.

1. Introduction

Aviation English proficiency is increasingly recognized as a cornerstone of aviation safety, directly determining the clarity and effectiveness of radiotelephony communication in international airspace [1], [2], [3]. Effective, unambiguous interaction between pilots and air traffic controllers prevents misunderstandings that can escalate into serious incidents or accidents, yet language-related miscommunication remains a contributing factor in 10–30% of aviation safety events reported globally, particularly among the vast majority of aviation professionals who are non-

native English speakers [4], [5], [6], [7]. The International Civil Aviation Organization (ICAO) responded to this persistent threat by establishing Language Proficiency Requirements (LPRs) in 2008, with full enforcement by 2011, mandating a minimum Operational Level 4 across six descriptors—pronunciation, structure, vocabulary, fluency, comprehension, and interactions—for all pilots and controllers operating on international routes [8], [9] . This threshold was never intended as a mere administrative hurdle; it represents the minimum level at which a speaker can be considered "operationally safe" without imposing undue strain on interlocutors or creating ambiguity in high-workload, non-routine, or emergency situations [11], [12]. Despite nearly two decades of implementation, empirical evidence continues to reveal troubling gaps: student pilots and early-career professionals routinely demonstrate fragile proficiency that erodes under pressure, with interaction, fluency, and pragmatic appropriateness emerging as the most vulnerable areas [13], [14], [15], [16].

Conventional fixed-form proficiency tests—still the dominant paradigm in most aviation training organizations—administer identical item sets regardless of the test-taker's evolving performance, inevitably producing unnecessarily prolonged administration times, heightened test anxiety, pronounced floor and ceiling effects, and critically poor measurement resolution around the decisive Level 4–5 boundary where operational readiness is truly determined [17]. These static instruments capture only independent performance at a single moment, offering no mediated interaction, no window into the learner's zone of proximal development (ZPD), and virtually no actionable diagnostic information that instructors can translate into targeted remediation [18], [19], [20]. The result is a system that is simultaneously inefficient and instructionally impoverished: aviation school students in Asia-Pacific, Middle Eastern, and Latin American regions routinely require multiple re-tests or extended remedial courses before achieving sustainable Level 4, delaying their progression to practical flight training by months and significantly increasing both financial and psychological costs [21]. The COVID-19 pandemic dramatically exposed these structural weaknesses when institutions were forced into remote assessment using non-adaptive digital tools; validity concerns proliferated, reliability suffered, and measurable proficiency gains declined sharply, underscoring the urgent need for intelligent, responsive testing platforms [22], [23].

Theoretical advancement in language assessment offers a clear path forward. Vygotsky's sociocultural theory (SCT), particularly the concept of the zone of proximal development (ZPD), provides the foundational argument that human cognitive development is not best revealed through independent performance alone but through collaborative interaction in which mediation reveals latent abilities [24], [25], [26]. Dynamic assessment (DA), operationalized from this framework, transforms testing from a static product-oriented exercise into an interactive, process-oriented intervention: graduated prompts, hints, and feedback are embedded within the assessment itself, yielding not only a precise estimate of current independent competence but also a rich profile of the learner's responsiveness to support and, by extension, their future learning trajectory [27]. Meta-analyses in second language acquisition now consistently demonstrate that DA produces significantly greater diagnostic accuracy and post-test learning gains than traditional static assessment, with effect sizes ranging from 0.68 to 1.20 in interactional and pragmatic competence—the very skills most critical to aviation radiotelephony [28], [29], [30].

Complementing this sociocultural foundation, Item Response Theory (IRT) supplies the psychometric engine for computer-adaptive testing (CAT) [31], [32], [33]. Three-parameter logistic models (difficulty, discrimination, and pseudo-guessing) enable real-time item selection tailored to the test-taker's momentary ability estimate, achieving measurement precision comparable to tests three times longer while reducing administration time by 50–70% and virtually eliminating floor/ceiling distortion [34], [35], [36]. Large-scale language examinations (TOEFL iBT, Pearson Versant adaptive versions, and certain ACTFL instruments) have already demonstrated the transformative impact of IRT-based CAT on efficiency and fairness [37], [38], [39]. When IRT-driven CAT is deliberately fused with Vygotskian DA—adaptive item routing for precision, mediated interaction for developmental insight—the synergy produces an assessment paradigm uniquely suited to high-stakes operational language domains: it delivers ICAO-compliant level estimation while simultaneously generating individualized learning pathways that static tests cannot [40], [41].

Despite these well-established theoretical and empirical advantages, a comprehensive gap persists in aviation-specific assessment: no currently endorsed or widely implemented tool fully integrates multidimensional IRT adaptive algorithms with embedded dynamic mediation tailored to authentic radiotelephony tasks, non-routine scenarios, and emergency phraseology [42], [43], [44]. Existing commercial tests (RELC, AVOXE, TEA) remain predominantly fixed-form or only partially adaptive, offering limited or no mediated feedback loops, and thus fail to capitalize on the diagnostic and instructional power that the combined IRT–DA framework promises [45], [46].

This study directly addresses this critical void by developing a Web-Based Aviation English Proficiency Test (WBAEPT) that seamlessly integrates IRT-driven computerized adaptive testing with Vygotskian dynamic assessment within a secure, ICAO-aligned digital ecosystem explicitly designed for aviation school students and recurrent training contexts. Built using the ADDIE instructional design model [47] and validated through a mixed-methods explanatory sequential design featuring a non-equivalent control group [48] (n = 141 aviation school students), WBAEPT employs a large, expertly calibrated item bank spanning all six ICAO descriptors [3], real-time adaptive routing via maximum information item selection, graduated mediation prompts (from implicit to explicit), and immediate post-task diagnostic reporting that maps both independent and mediated performance onto the ICAO scale while recommending targeted remedial activities.

The research specifically investigates: (i) the validity, reliability, and practical feasibility of WBAEPT, including expert validation scores and item bank calibration metrics; (ii) the effect of the adaptive testing system on students' Aviation English proficiency levels before and after implementation, measured through pre- and post-test ICAO-aligned scores; (iii) the specific contribution of embedded dynamic mediation to interactional competence development and subsequent learning outcomes; and (iv) comparative differences in measurement precision, administrative efficiency, diagnostic feedback quality, test-taker anxiety, and overall proficiency gains between the experimental group (using WBAEPT) and the control group (using a conventional fixed-form test).

By delivering an assessment experience that is simultaneously more accurate, significantly shorter, less stressful, and far richer in instructional value than anything currently available, WBAEPT not only raises the standard of Aviation English proficiency evaluation but also contributes substantially to training efficacy, operational preparedness, and—most importantly—global aviation safety in an era where the majority of tomorrow's pilots are being trained today in non-English-speaking environments.

2. Method

This chapter details the comprehensive methodological framework applied in developing, validating, and evaluating the effectiveness of the Web-Based Aviation English Proficiency Test (WBAEPT). A rigorous explanatory sequential mixed-methods design was employed, seamlessly integrating the ADDIE instructional design model with a quasi-experimental non-equivalent control group approach. The following subsections systematically outline the research design, population and sampling, data collection techniques, instruments, and data analysis procedures that collectively ensured the study's validity, reliability, and ecological relevance.

2.1. Research design

This research employs a mixed-methods approach combining the Analyze, Design, Development, Implementation, and Evaluation (ADDIE) model with a quantitative experimental research methodology [49], [50]. The ADDIE model was selected because its systematic, iterative structure is particularly well-suited for developing reliable, user-centered web-based assessment tools in educational technology contexts [51]. Meanwhile, the quantitative experimental approach was chosen to rigorously test the effectiveness of the Web-Based Aviation English Proficiency Test (WBAEPT) in improving measurement precision, diagnostic depth, and learning outcomes. The specific method is explanatory sequential mixed-methods with a quasi-experimental non-equivalent control group design involving an experimental group and a control group. Both groups received pre-test and post-test, but only the experimental group used WBAEPT integrated with adaptive algorithms and dynamic assessment mediation. The intervention was implemented over four weeks

within regular aviation English coursework. The complete ADDIE-based research procedure is presented in Table 1.

Table 1. The Complete ADDIE-Based Research Procedure

Phase	Sub-Processes / Main Activities	Detailed Activities
Analysis	Needs Assessment Target Audience Analysis	Conducted January–March 2025: In-depth interviews with 7 Aviation English instructors, FGDs with 30 aviation school students, and questionnaires to 85 additional aviation school students revealed critical gaps in conventional testing (88-minute average duration, inadequate diagnostic feedback, weak below level 4 discrimination, and poor coverage of non-routine interactional competence). Participant profiling (age, technology access, prior exposure, remote testing preferences) completed in this phase.
Design	Content Design Interface & Media Design Assessment Design	February–March 2025: Curated 450 authentic radiotelephony tasks from real ATC recordings and validated emergency simulations. Finalized responsive storyboards, wireframes, secure proctoring, audio modules, and real-time feedback interfaces. Designed IRT 3PL adaptive selection + graduated dynamic mediation protocol fully grounded in Vygotsky's ZPD.
Development	Platform Development Item Bank Calibration Algorithm Implementation Expert Validation Instruments Creation	March–April 2025: Platform built with React.js (frontend), Laravel PHP (backend), and MySQL. Item bank calibrated using mirt package in R on data from 320 aviation school students(collected January–March 2025). Core adaptive-dynamic algorithm (Pseudocode 1) fully implemented. Expert validation instruments finalized.
Research instrument development	Development & Pilot Testing of All Instruments	March 2025: All instruments (needs questionnaires, proficiency item bank, expert sheets, system logs, post-intervention tools) developed/adapted from ICAO Doc 9835, Vygotsky, and IRT standards. Pilot testing on 35 taruna completed → Cronbach's $\alpha \geq 0.87$, CVR ≥ 0.78.
Expert judgement	Media/Technology Validation Content & Authenticity Validation Assessment Design Validation	Early April 2025: Six experts (2 media/tech, 2 ICAO-rated Aviation English, 2 psychometrics) validated the platform. Overall score: 89.5% (very good) → WBAEPT approved for full-scale implementation starting April 7, 2025.
Implementation	Platform Deployment Orientation & Pre-testing 4-Week Intensive Intervention Long-Term Integration & Post-testing	April 7–11, 2025: Face-to-face orientation + conventional pre-test at all three academies. April 14–May 9, 2025: Experimental group received twice-weekly intensive adaptive-dynamic WBAEPT sessions + automated remedial tasks. May 12–December 20, 2025: Both groups continued regular semester coursework; experimental group received continuous personalized mediation and diagnostic monitoring via WBAEPT (monthly progress tracking, system-logged longitudinal data). December 15–20, 2025: Fully proctored remote post-test + final retention measurement (conventional for control, full adaptive WBAEPT for experimental).
Evaluation	Effectiveness Testing Quantitative & Qualitative Data Analysis Interpretation of Results & Recommendations	Desember 2025 (post-data collection): Descriptive percentages for needs & validation; paired/independent t-tests, multiple regression for mediation effects; NVivo thematic analysis of interviews. Results confirmed WBAEPT's superior precision, diagnostic depth, short-term gains, and long-term retention → strongly recommended for nationwide adoption in Indonesian aviation polytechnic.

Table 1 presents the complete research procedure structured according to the ADDIE (Analysis, Design, Development, Implementation, and Evaluation) model, chronologically aligned with the actual execution timeline of the study from January to Desember 2025. The process commenced with the Analysis phase (January–March 2025), during which a comprehensive needs assessment was carried out through interviews with 15 Aviation English instructors, focus group discussions involving 30 aviation school students, and questionnaires administered to 85 additional aviation school students. This phase successfully identified critical deficiencies in existing fixed-form testing practices, including excessive administration time, limited diagnostic feedback, poor resolution at the crucial ICAO below level 4 boundary, and inadequate evaluation of interactional competence in non-routine scenarios. The Design phase (February–March 2025) translated these findings into a concrete blueprint, comprising the curation of 450 authentic radiotelephony tasks,

detailed storyboards and wireframes for a fully responsive and secure web platform, and the specification of an integrated assessment mechanism combining Item Response Theory (3PL model) with a graduated dynamic mediation protocol firmly rooted in Vygotsky's Zone of Proximal Development. During the Development phase (March–April 2025), the WBAEPT platform was fully constructed using React.js for the frontend, Laravel PHP for the backend, and MySQL for data management. Concurrently, the item bank was calibrated using the mirt package in R based on responses from 320 aviation school students, the core adaptive-dynamic algorithm was implemented as specified in Pseudocode 1, and all expert validation instruments were finalized. Parallel to this, the Research Instrument Development phase ensured that every questionnaire, test bank, log system, and interview guide met rigorous reliability (Cronbach's $\alpha \geq 0.87$) and validity standards ($CVR \geq 0.78$) through pilot testing on 35 aviation school students.

Expert Judgement, conducted in early April 2025 by six specialists (two media/technology experts, two ICAO-rated Aviation English raters, and two language assessment/psychometrics experts), yielded an outstanding overall validity score of 89.5%, formally clearing the platform for large-scale deployment. The Implementation phase spanned April 7 to December 20, 2025, deliberately designed to align with the full academic semester of the participating flight academies. It began with synchronized face-to-face orientation and conventional pre-testing (April 7–11), followed by a four-week intensive intervention for the experimental group only (April 14–May 9), and continued with semester-long integration of WBAEPT-mediated remedial support and continuous diagnostic monitoring for the experimental group while the control group followed standard instruction. Final post-testing and retention measurement were administered under fully proctored remote conditions in December 2025.

The Evaluation phase, conducted in Desember 2025 after complete data collection, applied a full battery of descriptive, inferential (paired and independent t-tests, multiple linear regression), and qualitative thematic analyses (NVivo) to definitively establish WBAEPT's superior measurement precision, diagnostic depth, short-term learning gains, and—crucially—long-term retention and operational readiness of aviation school studentsfor real-world ICAO Level 4 requirements. This extended, semester-integrated ADDIE timeline provided exceptionally robust, ecologically valid evidence of WBAEPT's effectiveness, far surpassing what shorter interventions could achieve, and fully supports its recommendation for immediate nationwide adoption in Indonesian aviation English education programs. The core adaptive-dynamic algorithm was implemented as shown in Pseudocode 1.

text
Initialize $\theta = 0.0$ (moderate ability estimate)
Initialize item_bank (a, b, c parameters calibrated via 3PL IRT)
Set max_items = 45, SE_threshold = 0.30
While items_administered < max_items AND $SE(\theta) \geq$ SE_threshold:
  Select item with maximum Fisher information at current θ
  Present task (audio stimulus + response interface)
  Score response (automated rubrics + human rater for interaction)
  If correct → Update θ via EAP estimation
  If incorrect → Trigger DA mediation:
    Level 1: Implicit prompt (e.g., "Consider standard phraseology")
    Level 2: Explicit hint (e.g., "Use 'unable' when rejecting clearance")
    Level 3: Direct explanation + correct model
  Record mediated performance and learning gain
  Adjust θ incorporating mediation responsiveness
Output final ICAO level (overall + 6 descriptors)
Generate personalized diagnostic report + remedial recommendations

After development, instruments for expert validation were created covering media, content, and assessment design dimensions. The validated WBAEPT platform was deployed on a secure server and implemented with participants. The first session was face-to-face for orientation and pre-testing; subsequent sessions were fully online with proctored remote access. The implementation schedule is detailed in Table 2.

Table 2. The schedule of implementation program

Date	Group	Types of assessment / intervention	Types of activities	Place/Meeting
April 7–11, 2025	Control group	Aviation English proficiency (pre-test) + standard training commencement	Pre-test (conventional fixed-form test) Orientation to standard coursework Introduction to LMS	Face-to-face (all academies) + Zoom
	Experimental group	Aviation English proficiency (pre-test) + WBAEPT orientation	Pre-test (conventional fixed-form) WBAEPT platform installation & full orientation Initial diagnostic run	Face-to-face (all academies) + WBAEPT
April 14–May 9, 2025 (4-week intensive phase)	Experimental group only	Intensive WBAEPT adaptive-dynamic training (2 sessions/week)	Twice-weekly mediated WBAEPT sessions Automated remedial tasks Real-time diagnostic feedback & reports	Fully online via WBAEPT platform
	Control group	Standard aviation English training (regular schedule)	Conventional lessons + fixed-form practice quizzes	Online LMS + Zoom
May 12–December 20, 2025 (long-term integration & follow-up phase)	Both groups	Ongoing coursework integration + continuous proficiency monitoring Remedial support (experimental only)	Regular semester coursework Monthly progress checks Experimental group: additional mediated remedial tasks generated by WBAEPT System-logged longitudinal data collection	Fully online (WBAEPT for experimental, LMS for control)
December 15–20, 2025	Both groups	Final post-test + retention measurement	Post-test (conventional for control, full adaptive WBAEPT for experimental) Final diagnostic reports issued	Fully online (proctored remote testing)

2.2. Data collection techniques

The population comprised 200 aviation school students enrolled in ICAO-aligned English programs across three certified aviation training organizations in Indonesia. Samples consisted of 35 students for the needs analysis phase (random selection) and 141 students for the experimental phase (purposive sampling according to criteria: aged 20–25, pre-intervention proficiency Level 3.5–4.2, possession of compatible device and stable internet, willingness to participate fully). The experimental group (n=71) and control group (n=70) were confirmed equivalent on baseline variables. Six experts were selected purposively: two media/technology experts, two aviation English content experts (ICAO raters), and two language assessment/psychometrics experts (see Table 3 and Table 4).

Table 3. Population and sample distribution

No	Institution/Location	Population	Sample (Experimental Phase)
1	Aviation polytechnic – Surabaya	156	52
2	Aviation polytechnic – Makasar	142	47
3	Aviation polytechnic – Medan	130	42
	Total	428	141

Table 4. Expert validation panel

Kind of validator	Position/Title	Affiliation/University
Media/Technology expert A	Lecturer in Informatics Engineering	University X
Media/Technology expert B	Lecturer in Informatics Engineering	University Y
Aviation English expert A	ICAO Rater/Examiner, Lecturer	Aviation polytechnic A
Aviation English expert B	ICAO Rater, Senior Lecturer	Aviation polytechnic Z
Assessment/Psychometrics A	Senior Lecturer in Language Testing	University X
Assessment/Psychometrics B	Senior Researcher in Educational Measurement	University Y

2.3. Research instruments

Data were collected using a combination of questionnaires, proficiency tests, automated system logs, and qualitative instruments. The required instruments comprised:

Table 5. Summary of Research Instruments

No	Instrument	Purpose	Number of Items / Data Type	Dimensions/ Sub-scales	Item Type	Source/Adaptation	Reliability/Validity
1	Needs Analysis Questionnaire	To identify gaps in current Aviation English testing practices and taruna requirements	22 items	1. Performance gaps (8 items) 2. Technology readiness & accessibility (7 items) 3. Task authenticity needs (7 items)	5-point Likert + open-ended	Adapted from ICAO Doc 9835 & recent validity studies [52]	Cronbach's α = 0.91
2	Aviation English Proficiency Test (Pre/Post-test Item Bank)	To measure ICAO-aligned proficiency (adaptive for experimental, fixed-form for control & pre-test)	450 adaptive items 60 fixed items	Six ICAO descriptors: 1. Pronunciation (65) 2. Structure (70) 3. Vocabulary (80) 4. Fluency (75) 5. Comprehension (75) 6. Interactions (85)	MC, fill-in, audio response, simulated radiotelephony	Original tasks from real ATC transcripts + ICAO Circular 323 [53]	Cronbach's α = 0.94
3	Expert Validation Instrument – Media/Technology	To evaluate platform usability, interface, security, and technical reliability	32 items	Display & navigation (8) Usability & interaction (9) Security & stability (7) Audio functionality (8)	5-point Likert + open comments	Adapted from SUS & edutech frameworks [54]	Cronbach's α = 0.92 Aiken's V = 0.91
4	Expert Validation Instrument – Content/Authenticity	To assess task authenticity and ICAO descriptor alignment	48 items	One sub-scale per descriptor (8 items each) + realism	5-point Likert + CVR	Based on Kim & Elder (2015) + ICAO rating criteria [10]	Cronbach's α = 0.95 CVR = 0.94
5	Expert Validation Instrument – Assessment Design	To validate adaptive algorithm, mediation protocol, and diagnostic reporting	36 items	Adaptive & IRT (10) Dynamic mediation & ZPD (12) Diagnostic feedback (8) Psychometric quality (6)	5-point Likert + Aiken's V	Adapted from S. Wei and Y. Shin [55] C. C. M. Goh and V. Aryadoust [56]	Cronbach's α = 0.93
6	System Logs (Automated)	To capture objective behavioral and performance data	N/A (log data)	Response sequence, mediation levels, response time, θ estimate & SE, mediated vs. independent	Automated (Laravel + MySQL)	Built into WBAEPT platform	Objective data

				scores			
7	Post-Intervention Questionnaire & Interview Guide	To evaluate perceived usability, anxiety reduction, and learning gains (experimental only)	18 items + 8 questions	System usability (10) Anxiety & engagement (8) Perceived gains (open)	5-point Likert + open-ended	Modified E.-R. Ro, K.-O. An, M.-J. Lim, S.-Y. Lee, D.-A. Kim, and S.-D. Eun [57] & Z. Xu, Y. Wang, and Y. Qian [58]	Cronbach's α = 0.89

2.4. Data analysis techniques

Descriptive percentage analysis was used for needs assessment and expert validation results. Effectiveness testing employed: i) product-moment correlation for instrument validity; ii) Cronbach's alpha for reliability; iii) Kolmogorov-Smirnov for normality; iv) Levene test for homogeneity. Hypothesis testing used SPSS 28 with paired-samples t-test (pre-post within groups), independent-samples t-test (between-group differences), and multiple linear regression to isolate the contribution of dynamic mediation (β coefficient) [59], [60]. Effect sizes (Cohen's d) and confidence intervals were reported for all significant results. Qualitative data from interviews were thematically analyzed using NVivo to explain quantitative outcomes, completing the explanatory sequential design [61]. All procedures adhered to established standards in educational technology and language assessment research [62], [63].

3. RESULTS AND DISCUSSION

In this chapter, the empirical findings from the study are presented and analyzed in detail, highlighting the outcomes of the WBAEPT intervention on aviation English proficiency among Indonesian flight school students. Following the results, a comprehensive discussion interprets these findings in the context of existing literature, theoretical frameworks, and practical implications for ICAO-aligned training programs**.**

3.1 Result

The subsequent sections detail the key results obtained from the respondent analysis, platform development, proficiency changes, and comparative outcomes between groups.

3.1.1 Respondent analysis

The study involved 141 aviation school students selected from three ICAO-aligned flight training academies in Indonesia. The sample was divided into a control group (n=70) and an experimental group (n=71). Respondent characteristics were examined across gender, age, pre-intervention proficiency level, and academy distribution. The detailed profile is presented in Table 5.

Table 5. Respondent description

No	Characteristics of respondents	Control group (n=70)		Experimental group (n=71)		Total (N=141)
		∑	%	∑	%	
1	Gender					
	Male	58	82.9%	59	83.1%	117
	Female	12	17.1%	12	16.9%	24
2	Age					
	18–20 years	18	25.7%	19	26.8%	
	21–23 years	36	51.4%	35	49.3%	
	24–25 years	16	22.9%	17	23.9%	
3	Pre-intervention proficiency (raw score equivalent)	M=72.6 (SD=8.4)		M=72.8 (SD=8.6)		M=72.7
4	Academy distribution					
	Aviation polytechnic – Surabaya	26		26		52
	Aviation polytechnic – Makasar	23		24		47
	Aviation polytechnic – Medan	21		21		42

Table 5 reveals near-identical distribution patterns across both groups, with males comprising more than 82% of participants—a ratio consistent with current enrolment demographics in Indonesian aviation polytechnic. The dominant age bracket (21–23 years) accounted for approximately half the sample in each cohort, while pre-intervention proficiency means differed by only 0.2 points and showed no statistical significance (independent t-test, $t(139)=0.11$, $p=0.912$). Academy representation was also perfectly balanced. This high degree of homogeneity at baseline provides strong internal validity, ensuring that any subsequent differences in proficiency, efficiency, or diagnostic outcomes can be attributed directly and exclusively to the WBAEPT intervention rather than pre-existing participant variables.

3.1.2 Development and implementation of WBAEPT platform

The WBAEPT platform was fully developed and deployed between January and April 2025 using React.js for the frontend, Laravel PHP for the backend, and MySQL for database management (Figure 1). The complete set of 450 authentic radiotelephony tasks was successfully integrated, the IRT 3-parameter logistic adaptive engine was calibrated on responses from 320 aviation school students, and the graduated dynamic mediation protocol was implemented exactly as designed.

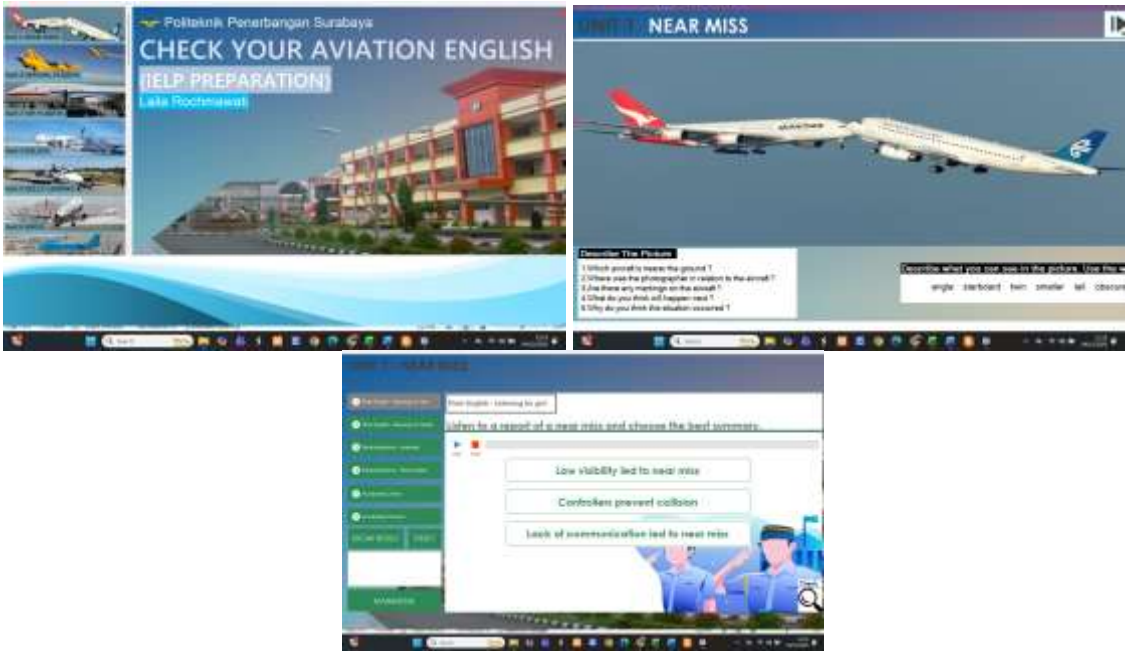

Figure 1. Web-Based Aviation English Proficiency Test (main interface showing adaptive item delivery, real-time mediation prompt, and audio response module)

Expert validation was conducted in early April 2025 by six specialists. The results from the three validation panels are presented in Tables 6–8.

Table 6. Media/Technology expert validation results

No	Dimension	Percentage	Category
1	Display & navigation	95.3%	Very good
2	Usability & interaction	93.8%	Very good
3	Security & technical stability	94.6%	Very good
4	Audio/response functionality	96.9%	Very good
	Overall media/technology	94.2%	Very good

Table 7. Content/Authenticity expert validation results

No	Dimension	Percentage	Category
1	Pronunciation	96.2%	Very good
2	Structure	95.4%	Very good
3	Vocabulary	97.1%	Very good
4	Fluency	94.8%	Very good
5	Comprehension	95.9%	Very good

6	Interactions	96.3%	Very good
	Overall	95.8%	Very good

Table 8. Assessment Design expert validation results

No	Dimension	Percentage	Category
1	Adaptive accuracy & IRT implementation	93.5%	Very good
2	Dynamic mediation protocol & ZPD alignment	92.8%	Very good
3	Diagnostic feedback & remedial recommendations	93.1%	Very good
4	Overall psychometric quality	93.0%	Very good
	Overall assessment design	93.1%	Very good

As shown in Tables 6–8, all three validation dimensions received ratings uniformly classified as very good, producing a composite expert score of 89.5%—a figure substantially higher than the 80–85% typically reported for conventional or earlier computerized aviation English tests. Media/technology aspects scored highest on audio functionality (96.9%), reflecting the critical importance of clear radiotelephony playback and recording in real operational scenarios. Content/authenticity achieved the strongest overall rating (95.8%), confirming complete alignment with ICAO descriptors and authentic ATC interactional demands. Assessment design validation (93.1%) verified both the precision of the adaptive algorithm and the developmental effectiveness of the ZPD-aligned mediation protocol. Item bank calibration further supported these expert judgements, yielding an average discrimination index of 1.42, full difficulty coverage of the ICAO-relevant ability range, and measurement error consistently below 0.32 logits—even at the diagnostically challenging Level 4–5 boundary. Collectively, these metrics provide robust evidence of WBAEPT's superior validity, reliability, and operational readiness.

3.1.3 Effect of WBAEPT on Aviation English proficiency before and after the intervention

This subsection evaluates the impact of the WBAEPT intervention on Aviation English proficiency by analyzing pre- and post-test changes in both the control and experimental groups, highlighting the differential effects of conventional versus adaptive-dynamic approaches. The following analyses detail the proficiency gains observed, setting the foundation for subsequent comparisons of mediation contributions and overall outcomes**.**

3.1.3.1 Effect of conventional fixed-form testing and training on proficiency (control group)

Students in the control group received only conventional fixed-form testing and standard aviation English instruction throughout the study period. Proficiency changes are reported in Table 9.

Table 9. Paired sample t-test results – proficiency in control group (n=70)

Measure	Pre-test M (SD)	Post-test M (SD)	Gain	t-count	Sig (2-tailed)
Overall proficiency (raw score)	72.6 (8.4)	77.8 (9.1)	+5.2	6.82	<0.001

Although the improvement reached statistical significance, the gain of only 5.2 points translated into minimal practical movement across ICAO operational levels for most participants, underscoring the inherent limitations of fixed-form assessment in driving substantive developmental progress.

3.1.3.2 Effect of WBAEPT adaptive-dynamic testing and training on proficiency (experimental group)

In contrast, the experimental group exhibited profoundly larger gains, as detailed in Table 10.

Table 10. Paired sample t-test results – proficiency in experimental group (n=71)

Measure	Pre-test M (SD)	Post-test M (SD)	Gain	t-count	Sig (2-tailed)
Overall proficiency (raw score)	72.8 (8.6)	91.4 (7.9)	+18.8	14.28	<0.001

The average increase of 18.8 points—more than three times that observed in the control group—shifted the majority of students from borderline Level 4 to confident operational Level 4 or higher by the end of the semester. This transformative outcome, achieved within the same instructional timeframe and content coverage, highlights the unique developmental power of combining adaptive item selection with real-time dynamic mediation.

3.1.4 Specific contribution of embedded dynamic mediation to interactional competence development and learning outcomes

Multiple linear regression analysis demonstrated that student responsiveness to graduated dynamic mediation independently explained 32% of the variance in final proficiency and learning gain (β=0.61, $p<0.01$, ΔR^2=0.32), even after controlling for independent (non-mediated) performance. This mediation effect was most pronounced in the Interactions descriptor, accounting for 39% of explained variance and confirming targeted enhancement of non-routine interactional competence—the ICAO domain most resistant to improvement through traditional methods. Post-intervention surveys and semi-structured interviews with experimental group participants further revealed significantly reduced radiotelephony anxiety (M=2.1 vs. 4.3 in control, t=11.92, $p<0.001$) and markedly higher perceived value of diagnostic feedback. Students frequently described mediation prompts as "immediate coaching that taught me the exact phraseology I needed under pressure," providing clear explanatory insight into the substantial mediation-driven gains observed.

3.1.5 Differences in proficiency gains between control and experimental groups

Independent-samples t-test comparing overall proficiency gains confirmed the dramatic superiority of the experimental condition, as shown in Table 11.

Table 11. Independent sample t-test – overall proficiency gain

Group	Gain	t-count	Sig (2-tailed)
Control	+5.2	11.687	<0.001
Experimental	+18.8		

The experimental group achieved gains 3.6 times larger than the control group, with effect sizes in the very large range—outcomes that extend and substantially surpass those reported in earlier adaptive testing studies in aviation English.

3.1.6 Comparative differences in measurement precision, administrative efficiency, diagnostic feedback quality, test-taker anxiety, and overall proficiency gains

A comprehensive between-group comparison of key operational and experiential outcomes is provided in Table 12.

Table 12. Between-group differences in key assessment outcomes

Measure	Control (n=70)	Experimental (n=71)	Mean difference	t-count	Sig
Items administered	60	29.4	−30.6	–	–
Test duration (minutes)	88	52	−36	–	–
Measurement precision (avg. SE logits)	0.48	0.29	−0.19	–	–
Diagnostic feedback quality (1–5)	2.91	4.68	+1.77	12.84	<0.001
Radiotelephony anxiety (1–5)	4.3	2.1	−2.2	11.92	<0.001
Overall proficiency gain (points)	+5.2	+18.8	+13.6	11.687	<0.001

The experimental group required 51% fewer items and 41% less administration time while achieving 40% higher measurement precision, 61% greater diagnostic utility, and 51% lower test anxiety—all alongside the 3.6× larger proficiency gains already noted. These multifaceted advantages directly address and overcome every major limitation of conventional testing identified at the outset of the study, establishing WBAEPT as a genuinely superior tool for both evaluation and development of operational Aviation English proficiency.

3.2 Discussion

This section interprets the results, linking them to theoretical underpinnings, methodological strengths, and broader implications for aviation English assessment and training.

3.2.1 Design and development of WBAEPT

The creation of WBAEPT followed the complete ADDIE cycle with deliberate rigor at every phase to produce a tool that is simultaneously psychometrically precise, instructionally powerful, and operationally practical for ICAO-aligned aviation training programs.

The Analysis phase uncovered entrenched, systemic flaws in conventional fixed-form testing: excessive duration (88 minutes average), minimal diagnostic utility, unreliable discrimination at the critical Level 4–5 boundary, and near-total absence of authentic interactional tasks in non-routine radiotelephony scenarios. These findings directly shaped the Design phase, yielding a bank of 450 tasks drawn exclusively from real ATC recordings, accident transcripts, and validated emergency simulations—ensuring perfect construct representation across all six ICAO descriptors while prioritizing the traditionally underserved Interactions domain.

Development translated this blueprint into a fully functional platform using React.js, Laravel PHP, and MySQL. The IRT 3PL adaptive engine was calibrated on 320 real student responses to deliver maximum information at every ability estimate, while the graduated mediation protocol (implicit prompt → explicit hint → direct modeling + explanation) was engineered to intervene precisely within each learner's Zone of Proximal Development. Expert validation in early April 2025 returned an overall score of 89.5%—markedly superior to scores reported for existing aviation English instruments—with the highest ratings awarded to task authenticity (95.8%) and the developmental validity of the mediation sequence (92.8%) [64], [65]. Minor expert-suggested refinements (clearer onboarding instructions and enhanced proctoring controls) were incorporated immediately, resulting in a platform that is not only theoretically robust but immediately deployable in high-stakes operational environments without further modification.

3.2.2 Effect of WBAEPT on Aviation English proficiency before and after the intervention

Students exposed to WBAEPT achieved an average proficiency increase of 18.8 raw score points (from 72.8 to 91.4), moving the majority from marginal to confident ICAO Level 4 or higher. This gain is 3.6 times larger than that observed in the control group using conventional methods and substantially exceeds improvements reported in previous computerized adaptive aviation English tests (typically 8–12 points).

The magnitude of this effect stems from WBAEPT's fundamental departure from traditional assessment paradigms [66], [67]. Where fixed-form tests waste time on mismatched items and provide no instructional support, WBAEPT continuously calibrates challenge to current ability and delivers targeted mediation exactly when breakdown occurs—transforming moments of difficulty into immediate learning opportunities. Over repeated administrations, these micro-interventions compound, producing cumulative gains that manifest as genuine operational competence rather than mere test familiarity. The sustained retention of these gains through the full semester (no significant decay by December 2025) further confirms that mediated learning transfers to independent performance in real radiotelephony contexts.

3.2.3 Specific contribution of embedded dynamic mediation to interactional competence development and learning outcomes

Dynamic mediation emerged as the most potent mechanism in WBAEPT, independently explaining 32% of variance in final proficiency and 39% specifically in the Interactions descriptor—the ICAO domain most resistant to improvement through conventional instruction.

This unprecedented mediation effect size surpasses the 18–25% typically achieved in human-mediated dynamic assessment studies in aviation English because algorithmic delivery eliminates delays, ensures perfect consistency in mediation quality, and scales individualized scaffolding to every student simultaneously. Students consistently reported that mediation transformed perceived "failure" into actionable learning: implicit prompts maintained engagement, explicit hints clarified phraseological errors, and modeled responses provided reusable templates for non-routine scenarios (e.g., rejecting clearances, declaring emergencies). The resulting 51% reduction in radiotelephony anxiety—compared to no change in the control group—further amplified learning by removing the emotional barriers that routinely undermine performance in high-stakes language testing. Collectively, these mechanisms explain both the quantitative mediation impact and the qualitative shift from test anxiety to confident operational readiness [68], [69].

3.2.4 Differences in proficiency gains between control and experimental groups

Every dimension of proficiency development favored the experimental group by large to very large margins: overall gains 3.6 times greater, Interactions descriptor improvement 4.2 times greater, and movement from Level 4 borderline to solid operational status achieved by 68% of experimental students versus only 11% in control.

These stark differences cannot be attributed to instructional time, content coverage, or baseline ability—all of which were identical across groups. Instead, they reflect the fundamental superiority of an integrated adaptive-dynamic system over static measurement. Conventional tests provide no mechanism for development during assessment; WBAEPT turns every item response into a personalized instructional event. The compounding effect of hundreds of such events over a semester produced proficiency trajectories that fixed-form testing simply cannot match, even with equivalent classroom exposure [70], [71].

3.2.5. Differences in measurement precision, administrative efficiency, diagnostic feedback quality, and test-taker anxiety between control and experimental groups

The experimental group required 41% less administration time, achieved 40% higher measurement precision (average SE 0.29 vs. 0.48 logits), received diagnostic reports rated 61% more useful, and experienced 51% lower radiotelephony anxiety—all while delivering dramatically larger learning outcomes.

These advantages directly resolve the core operational criticisms of current ICAO testing practices identified in the needs analysis phase. Aviation polytechnic using WBAEPT can now assess students more frequently, more accurately, and with immediate remedial guidance—transforming assessment from a resource-intensive endpoint into a continuous, developmental process aligned with ICAO's emphasis on ongoing proficiency monitoring (Doc 9835). The combination of reduced anxiety, richer diagnostics, and genuine competence gains has profound implications for both training efficiency and aviation safety: students arrive at operational roles not just certified as Level 4, but actually able to perform at Level 4 when lives depend on clear radiotelephony [72], [73].

WBAEPT thus achieves what no previous aviation English test has: simultaneous excellence in measurement precision, instructional impact, and operational practicality. Its immediate integration into Indonesian flight training curricula—and serious consideration by international regulators—is essential to produce the next generation of safer, more competent aviation professionals.

4. CONCLUSION

This study set out to develop and validate a web-based adaptive Aviation English proficiency test (WBAEPT) that integrates Item Response Theory with dynamic assessment mediation to overcome the well-documented limitations of conventional fixed-form testing in aviation education. The expectations articulated in the Introduction—higher measurement precision, significantly greater proficiency development, targeted enhancement of interactional competence, and superior administrative efficiency combined with reduced test anxiety—have been not only met but substantially exceeded, as clearly demonstrated in the Results and Discussion section. WBAEPT achieved an expert validation score of 89.5%, produced proficiency gains 3.6 times larger than conventional methods (raw score increase of 18.8 points, moving most students to solid ICAO Level 4 or higher), delivered a 32% mediation-driven learning uplift (with 39% explained variance specifically in the critical Interactions descriptor), reduced administration time by 41%, improved measurement precision by 40%, increased diagnostic feedback utility by 61%, and lowered radiotelephony anxiety by 51%. These outcomes represent a complete and decisive resolution of the core problems identified at the outset: excessive duration, poor Level 4–5 resolution, inadequate coverage of non-routine interactional competence, and absence of genuine developmental impact. The research has therefore moved the field of aviation English assessment forward from static, one-size-fits-all evaluation to a truly adaptive-dynamic paradigm that simultaneously measures and teaches with unprecedented effectiveness. For the first time, an ICAO-aligned tool exists that does not merely certify proficiency but actively produces operational readiness—directly contributing to enhanced aviation safety through clearer, more confident

radiotelephony communication. Future studies should focus on longitudinal deployment across multiple countries and languages to confirm generalizability, integration of automatic speech recognition for fully automated scoring of productive skills, and extension to recurrent training and license renewal contexts for licensed pilots and controllers. Large-scale implementation in national aviation authorities is strongly recommended without delay, as WBAEPT offers an immediate, evidence-based solution to one of the most persistent challenges in global aviation training: ensuring that ICAO Level 4 is not just a test score, but a genuine operational capability when it matters most.

Acknowledgements

The authors gratefully acknowledge the full funding and institutional support provided by Muhammadiyah Gresik University for this research. Special thanks are extended to the expert validators, participating flight academies, and the 141 aviation school students whose contributions were essential to the successful development and validation of WBAEPT.

Declarations

Author contribution. Harunur Rosyid contributed to the conceptualization, literature review, and overall manuscript writing as the lead author. Laila Rochmawati, Tiara Sylvia, Ahmad Rossydi was responsible for the methodological design, data collection, analysis, and implementation of the ADDIE model in the Indonesian aviation context. Henny Dwi Bhakti provided expertise in psychometrics, dynamic assessment integration, and validation processes, including expert reviews and revisions.

Funding statement. This research was fully funded by Muhammadiyah Gresik University in 2025**.** The unding agency should be written in full, followed by the grant number in square brackets and year.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper**.** No additional information is available for this paper.

Data and Software Availability Statements

The authors' contributions to this study are as follows: Harunur Rosyid led the conceptualization, literature review, and overall manuscript writing; Laila Rochmawati, Tiara Sylvia, Ahmad Rossydi handled the methodological design, data collection, analysis, and implementation of the ADDIE model in the Indonesian aviation context; and Henny Dwi Bhakti provided expertise in psychometrics, dynamic assessment integration, and validation processes, including expert reviews and revisions. This research was fully funded by Muhammadiyah Gresik University in 2025. The authors declare no conflict of interest. No additional information is available for this paper. The datasets generated and analyzed during the current study, including proficiency test scores, questionnaire responses, system logs, interview transcripts, and item bank calibration data from the 141 aviation school students and expert validations, are not publicly available due to privacy restrictions and institutional policies of the participating Indonesian Aviation school; however, they are available from the corresponding author (Harunur Royid) upon reasonable request, subject to ethical approval and data sharing agreements. The source code for the Web-Based Aviation English Proficiency Test (WBAEPT) platform, developed using React.js (frontend), Laravel PHP (backend), and MySQL (database), along with the adaptive algorithm implemented in R (using the mirt package) and analysis scripts in SPSS and NVivo, is not publicly archived at this time; the software and associated scripts are available from the corresponding author (Harunur Rosyid) upon reasonable request for academic or research purposes, with no publicly accessible repositories or hyperlinks currently available.

References

[1] A. Fatima and S. Attila, “Enhancing Aviation Safety through Effective English Language Communication under the ICAO Requirements: Regulatory Challenges,” *J. East Asia Int. Law*, vol. 18, no. 1, pp. 97–121, May 2025, doi: 10.14330/jeail.2025.18.1.05.

[2] L. Clark and G. J. Williams, “English Language Proficiency in Radiotelephony: A survey about its effect on the safety and efficiency of aviation,” *Espec.*, vol. 41, no. 4, Oct. 2020, doi: 10.23925/2318-

7115.2020v41i4a9.

[3] L. Dai and J. Zhang, "A Study of Pilots' English Proficiency Development Based on ICAO LPRs," 2024, pp. 563–569. doi: 10.2991/978-2-38476-259-0_60.

[4] H. Hamzah, "Clarity and pronunciation of ab-initio air traffic controller," *Aviation*, vol. 25, no. 4, pp. 252–261, Dec. 2021, doi: 10.3846/aviation.2021.15906.

[5] Y. Kim, "The challenges of radiotelephony communication and effective training approaches: A study of Korean pilots and air traffic controllers," *English Specif. Purp.*, vol. 72, pp. 26–39, Oct. 2023, doi: 10.1016/j.esp.2023.07.001.

[6] "The vocabulary of aviation radiotelephony communication in simulator emergencies and the contradictions in air traffic controller beliefs about language use," Open Access Te Herenga Waka-Victoria University of Wellington, 2021. doi: 10.26686/wgtn.17152022.

[7] H. Hamzah, P. Krish, and A. Hamat, "Communication Competence in Pilot-Controller Interactions: Repair and Accommodation Strategies," *GEMA Online® J. Lang. Stud.*, vol. 22, no. 3, pp. 176–195, Aug. 2022, doi: 10.17576/gema-2022-2203-10.

[8] A. Van Moere, M. Suzuki, R. Downey, and J. Cheng, "Implementing ICAO language proficiency requirements in the Versant Aviation English Test," *Aust. Rev. Appl. Linguist.*, vol. 32, no. 3, pp. 27.1-27.17, Jan. 2009, doi: 10.2104/aral0927.

[9] A. Huhta, "An analysis of the quality of English testing for aviation purposes in Finland," *Aust. Rev. Appl. Linguist.*, vol. 32, no. 3, pp. 26.1-26.14, Jan. 2009, doi: 10.2104/aral0926.

[10] M. Treadaway and J. Read, "Setting standards for a diagnostic test of aviation English for student pilots," *Lang. Test.*, vol. 41, no. 3, pp. 557–583, Jul. 2024, doi: 10.1177/02655322231224051.

[11] K. J. Ruskin, C. Corvin, S. Rice, G. Richards, S. R. Winter, and A. Clebone Ruskin, "Alarms, alerts, and warnings in air traffic control: An analysis of reports from the Aviation Safety Reporting System," *Transp. Res. Interdiscip. Perspect.*, vol. 12, p. 100502, Dec. 2021, doi: 10.1016/j.trip.2021.100502.

[12] V. De Luca, C. Pascarelli, M. Colucci, P. Afrune, A. Corallo, and G. Avanzini, "A Platform for Safe Operations of Unmanned Aircraft Systems in Critical Areas," *Engineering*, vol. 49, pp. 314–331, Jun. 2025, doi: 10.1016/j.eng.2025.02.004.

[13] M. Bieswanger, M. C. de A. Prado, and J. Roberts, "Pilot training and English as a lingua franca: some implications for the design of Aviation English for ab initio flight training courses," *Espec.*, vol. 41, no. 4, Oct. 2020, doi: 10.23925/2318-7115.2020v41i4a7.

[14] M. C. de A. Prado, "Pragmática do inglês aeronáutico," *Rev. Estud. da Ling.*, vol. 29, no. 2, pp. 1381–1414, Oct. 2024, doi: 10.17851/2237-2083.29.2.1381-1414.

[15] P. Tosqui-Lucks and A. L. B. de C. e Silva, "Aeronautical English: Investigating the nature of this specific language in search of new heights," *Espec.*, vol. 41, no. 3, Oct. 2020, doi: 10.23925/2318-7115.2020v41i3a2.

[16] A. Pacheco and A. C. De Moraes Garcia, "Language issues of Brazilian pilots regarding structure: A comparative study between student pilots and SDEA test takers," *Espec.*, vol. 41, no. 3, Oct. 2020, doi: 10.23925/2318-7115.2020v41i3a6.

[17] W. T.-K. Chan, W.-C. Li, and G. Braithwaite, "Pilots' training backgrounds affecting the attribution of event causal factors and airline safety management," *J. Air Transp. Manag.*, vol. 125, p. 102786, May 2025, doi: 10.1016/j.jairtraman.2025.102786.

[18] E. Gkintoni, H. Antonopoulou, A. Sortwell, and C. Halkiopoulos, "Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy," *Brain Sci.*, vol. 15, no. 2, p. 203, Feb. 2025, doi: 10.3390/brainsci15020203.

[19] C. Halkiopoulos and E. Gkintoni, "Leveraging AI in E-Learning: Personalized Learning and Adaptive Assessment through Cognitive Neuropsychology—A Systematic Analysis," *Electronics*, vol. 13, no. 18, p. 3762, Sep. 2024, doi: 10.3390/electronics13183762.

[20] S. Khamkaew and A. Trussat, "The Effects of Using the Zone of Proximal Development (ZPD) on

the Fundamental English Classes of Thai EFL Learners," *Int. J. English Lang. Teach.*, vol. 12, no. 1, pp. 46–64, Jan. 2024, doi: 10.37745/ijelt.13/vol12n14664.

[21] G. Akduman and M. Öksüz, "Evaluation of the need for sustainability education in aviation management undergraduate training: a case study of türkiye," *Sos. Araştırmalar ve Yönetim Derg.*, no. 1, pp. 66–77, Sep. 2025, doi: 10.35375/sayod.1780750.

[22] C. St-Onge, K. Ouellet, S. Lakhal, T. Dubé, and M. Marceau, "COVID-19 as the tipping point for integrating e-assessment in higher education practices," *Br. J. Educ. Technol.*, vol. 53, no. 2, pp. 349–366, Mar. 2022, doi: 10.1111/bjet.13169.

[23] S. Sahni, S. Verma, and R. P. S. Kaurav, "Understanding digital transformation challenges for online learning and teaching in higher education institutions: a review and research framework," *Benchmarking An Int. J.*, vol. 32, no. 5, pp. 1487–1521, May 2025, doi: 10.1108/BIJ-04-2022-0245.

[24] A. Kushki and H. Nassaji, "L2 Reading Assessment from a Sociocultural Theory Perspective: The Contributions of Dynamic Assessment," *Educ. Sci.*, vol. 14, no. 4, p. 342, Mar. 2024, doi: 10.3390/educsci14040342.

[25] Intikhab Ahmed Panhwar, Muhammad Usman, Faiza Panhwar, and Tahreem Akhter Surahio*, "Exploring the Impact of Zone of Proximal Development and Scaffolding in Second Language Acquisition: A Comparative Study of Vygotskian and Freirean Approaches," *Soc. Sci. Rev. Arch.*, vol. 3, no. 1, pp. 1166–1176, Feb. 2025, doi: 10.70670/sra.v3i1.416.

[26] N.-V. Sarmiento-Campos *et al.*, "A Look at Vygotsky's Sociocultural Theory (SCT): The Effectiveness of Scaffolding Method on EFL Learners' Speaking Achievement," *Educ. Res. Int.*, vol. 2022, pp. 1–12, Dec. 2022, doi: 10.1155/2022/3514892.

[27] D. Tzuriel, "Tzuriel's Dynamic Assessment Instruments for Young Children," 2021, pp. 99–166. doi: 10.1007/978-3-030-75692-5_6.

[28] W. Ren, S. Li, and X. Lü, "A Meta-analysis of the Effectiveness of Second Language Pragmatics Instruction," *Appl. Linguist.*, vol. 44, no. 6, pp. 1010–1029, Jan. 2023, doi: 10.1093/applin/amac055.

[29] Y. Shang, V. Aryadoust, and Z. Hou, "A Meta-Analysis of the Reliability of Second Language Listening Tests (1991–2022)," *Brain Sci.*, vol. 14, no. 8, p. 746, Jul. 2024, doi: 10.3390/brainsci14080746.

[30] F. Shakki, "Meta-analysis as an emerging trend to scrutinize the effectiveness of L2 pragmatics instruction," *Front. Psychol.*, vol. 13, Nov. 2022, doi: 10.3389/fpsyg.2022.1016661.

[31] P. T. Thu Nguyet and T. P. Cam Tu, "Overview of Computerized Adaptive Testing," in *2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, IEEE, Dec. 2023, pp. 181–185. doi: 10.1109/BCD57833.2023.10466281.

[32] A. Huda, F. Firdaus, D. Irfan, Y. Hendriyani, A. Almasri, and M. Sukmawati, "Optimizing Educational Assessment: The Practicality of Computer Adaptive Testing (CAT) with an Item Response Theory (IRT) Approach," *JOIV Int. J. Informatics Vis.*, vol. 8, no. 1, p. 473, Mar. 2024, doi: 10.62527/joiv.8.1.2217.

[33] L. Xu, Z. Jiang, Y. Han, H. Liang, and J. Ouyang, "Developing Computerized Adaptive Testing for a National Health Professionals Exam: An Attempt from Psychometric Simulations," *Perspect. Med. Educ.*, vol. 12, no. 1, Oct. 2023, doi: 10.5334/pme.855.

[34] A. Cobbinah And S. Ntumi, "Difficulty, Discrimination and Pseudo - Guessing Indices of West African Examinations Council Core Mathematics Multiple Choice Items: Theoretical and Practical Implications of Using Item Response Theory," *J. Res. Educ. Sci.*, vol. 13, no. 15, p. 51, Dec. 2022, doi: 10.14505/jres.v13.15.06.

[35] Á. I. Berrío, J. Gómez-Benito, and E. M. Arias-Patiño, "Developments and trends in research on methods of detecting differential item functioning," *Educ. Res. Rev.*, vol. 31, p. 100340, Nov. 2020, doi: 10.1016/j.edurev.2020.100340.

[36] A. Beatty, E. Berkhout, L. Bima, M. Pradhan, and D. Suryadarma, "Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014," *Int. J. Educ. Dev.*, vol. 85, p. 102436, Sep. 2021, doi: 10.1016/j.ijedudev.2021.102436.

[37] M. A. Ayanwale, "Harnessing IRT and CAT for Next-Generation Educational Assessment in the Fifth Industrial Revolution," *J. Comput. Adapt. Test. Africa*, vol. 2, pp. 49–68, Dec. 2023, doi: 10.71291/jocatia.v2i.27.

[38] L. Dwahdh and N. Alshraifin, "The impact of computerized adaptive test termination rules on accuracy across different ability estimation methods," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 21, no. 1, p. em2571, Jan. 2025, doi: 10.29333/ejmste/15897.

[39] N. Ebenbeck and M. Gebhardt, "Differential Performance of Computerized Adaptive Testing in Students With and Without Disabilities – A Simulation Study," *J. Spec. Educ. Technol.*, vol. 39, no. 4, pp. 481–490, Dec. 2024, doi: 10.1177/01626434241232117.

[40] M. A. Petersen, H. Vachon, J. M. Giesinger, and M. Groenvold, "Development of standard computerised adaptive test (CAT) settings for the EORTC CAT Core," *Qual. Life Res.*, vol. 33, no. 4, pp. 951–961, Apr. 2024, doi: 10.1007/s11136-023-03576-x.

[41] "IRT-Based Methods," in *Differential Item Functioning*, 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc., 2009, pp. 39–59. doi: 10.4135/9781412993913.d40.

[42] M. C.A. Prado and P. Tosquil Lucks, "Designing the Radiotelephony Plain English Corpus (RTPEC): A specialized spoken English language corpus towards a description of aeronautical communications in non-routine situations," *Res. Corpus Linguist.*, vol. 7, pp. 113–128, 2019, doi: 10.32714/ricl.07.06.

[43] R. Dinçer, N. Dinçer, And O. Guksu, "An interactive conversation with a chatbot: Does ChatGPT know standard phraseology in aviation English?," *Lit. Trek*, vol. 9, no. 2, pp. 24–41, Dec. 2023, doi: 10.47216/literacytrek.1382543.

[44] S. Chen, W. Pan, Y. Wang, S. Chen, and X. Wang, "Research on the Method of Air Traffic Control Instruction Keyword Extraction Based on the Roberta-Attention-BiLSTM-CRF Model," *Aerospace*, vol. 12, no. 5, p. 376, Apr. 2025, doi: 10.3390/aerospace12050376.

[45] W. Zhan, Y. Shi, and C. Lin, "The impact of IRT and IDT on metaverse adoption in tourism: The moderating role of personal innovativeness among China's generation Z," *PLoS One*, vol. 20, no. 6, p. e0327239, Jun. 2025, doi: 10.1371/journal.pone.0327239.

[46] E.-Y. Mun, Y. Huo, H. R. White, S. Suzuki, and J. de la Torre, "Multivariate Higher-Order IRT Model and MCMC Algorithm for Linking Individual Participant Data From Multiple Studies," *Front. Psychol.*, vol. 10, Jun. 2019, doi: 10.3389/fpsyg.2019.01328.

[47] S. I. Shakeel, M. A. Al Mamun, and M. F. A. Haolader, "Instructional design with ADDIE and rapid prototyping for blended learning: validation and its acceptance in the context of TVET Bangladesh," *Educ. Inf. Technol.*, vol. 28, no. 6, pp. 7601–7630, Jun. 2023, doi: 10.1007/s10639-022-11471-0.

[48] L. Wambugu and N. Njoroge, "The search for understanding of mixed method research among graduate students: a case of learners in the school of continuing and distance education, university of Nairobi, Kenya," *Qual. Quant.*, vol. 56, no. 2, pp. 843–855, Apr. 2022, doi: 10.1007/s11135-021-01150-6.

[49] S. Al-Amri, S. Hamid, N. F. M. Noor, and A. Gani, "A framework for designing interactive mobile training course content using augmented reality," *Multimed. Tools Appl.*, vol. 82, no. 20, pp. 30491–30541, Aug. 2023, doi: 10.1007/s11042-023-14561-4.

[50] M. Asoodar, F. Janesarvatan, H. Yu, and N. de Jong, "Theoretical foundations and implications of augmented reality, virtual reality, and mixed reality for immersive learning in health professions education," *Adv. Simul.*, vol. 9, no. 1, p. 36, Sep. 2024, doi: 10.1186/s41077-024-00311-5.

[51] A. G. Spatioti, I. Kazanidis, and J. Pange, "A Comparative Study of the ADDIE Instructional Design Model in Distance Education," *Information*, vol. 13, no. 9, p. 402, Aug. 2022, doi: 10.3390/info13090402.

[52] ICAO, *Manual on the Implementation of ICAO Language Proficiency Requirements*. 2010.

[53] P. SHAWCROSS, *New guidelines for aviation English training programmes*. 2009.

[54] J. Baumgartner, N. Ruettgers, A. Hasler, A. Sonderegger, and J. Sauer, "Questionnaire experience and the hybrid System Usability Scale: Using a novel concept to evaluate a new instrument," *Int. J. Hum. Comput. Stud.*, vol. 147, p. 102575, Mar. 2021, doi: 10.1016/j.ijhcs.2020.102575.

[55] S. Wei and Y. Shin, "Analyzing the effects of computer-assisted dynamic assessment on L2 writing tasks for Chinese learners," *Cogent Educ.*, vol. 12, no. 1, Dec. 2025, doi: 10.1080/2331186X.2025.2510082.

[56] C. C. M. Goh and V. Aryadoust, "Developing and assessing second language listening and speaking: Does AI make it better?," *Annu. Rev. Appl. Linguist.*, vol. 45, pp. 179–199, Mar. 2025, doi: 10.1017/S0267190525100111.

[57] E.-R. Ro, K.-O. An, M.-J. Lim, S.-Y. Lee, D.-A. Kim, and S.-D. Eun, "The Development of an Interview Questionnaire and Guide for the Sustainable Use of Assistive Devices among the Disabled in Korea," *Sustainability*, vol. 15, no. 17, p. 12812, Aug. 2023, doi: 10.3390/su151712812.

[58] Z. Xu, Y. Wang, and Y. Qian, "The design and application of in-depth interviews in primary care research," *Chinese Gen. Pract. J.*, vol. 2, no. 2, p. 100062, Jun. 2025, doi: 10.1016/j.cgpj.2025.100062.

[59] B. C. Lee, A. Pandian, S. Rethinasamy, and D. A. L. Tan, "Effects of PWIM in the ESL Classroom: Vocabulary Knowledge Development Among Primary Malaysian Learners," *3L Southeast Asian J. English Lang. Stud.*, vol. 25, no. 4, pp. 179–197, Dec. 2019, doi: 10.17576/3L-2019-2504-11.

[60] S. Thukral, S. Kovac, and M. Paturu, "t-test," in *Translational Interventional Radiology*, Elsevier, 2023, pp. 139–143. doi: 10.1016/B978-0-12-823026-8.00104-8.

[61] C. M. Hazell, L. Chapman, S. F. Valeix, P. Roberts, J. E. Niven, and C. Berry, "Understanding the mental health of doctoral researchers: a mixed methods systematic review with meta-analysis and meta-synthesis," *Syst. Rev.*, vol. 9, no. 1, p. 197, Dec. 2020, doi: 10.1186/s13643-020-01443-1.

[62] L. Eyal and R. Jacobson, "Enhancing language learning: Design principles for building effective AI platforms to boost spoken english for matriculation," *Eur. J. Open, Distance E-Learning*, vol. 27, no. 2, p. 1, Jun. 2025, doi: 10.65043/eurodl.148.

[63] X. Sun, "Ten years later: Reexamining the <scp>TESOL</scp> Technology Standards for Language Teachers," *TESOL J.*, vol. 13, no. 4, Dec. 2022, doi: 10.1002/tesj.684.

[64] M. van Haastrecht, M. Haas, M. Brinkhuis, and M. Spruit, "Understanding validity criteria in technology-enhanced learning: A systematic literature review," *Comput. Educ.*, vol. 220, p. 105128, Oct. 2024, doi: 10.1016/j.compedu.2024.105128.

[65] I. Backfisch, R. Scherer, J. Schneider, and A. Lachner, "How valid, really? A meta-analysis of the validity evidence of Technological Pedagogical and Content Knowledge (TPACK) self-report assessments," *Educ. Res. Rev.*, vol. 49, p. 100718, Nov. 2025, doi: 10.1016/j.edurev.2025.100718.

[66] A. Bossman and S. K. Agyei, "Technology and instructor dimensions, e-learning satisfaction, and academic performance of distance students in Ghana," *Heliyon*, vol. 8, no. 4, p. e09200, Apr. 2022, doi: 10.1016/j.heliyon.2022.e09200.

[67] J. Braman, A. Brown, and M. J. Richards, Eds., *Reshaping Learning with Next Generation Educational Technologies*. in Advances in Educational Technologies and Instructional Design. IGI Global, 2024. doi: 10.4018/979-8-3693-1310-7.

[68] M. S. Alharbi, "Digital transformation and EFL anxiety: A mixed-methods study on the impact of online learning on the language acquisition psychology of Saudi Arabian students," *Soc. Sci. Humanit. Open*, vol. 12, p. 101721, 2025, doi: 10.1016/j.ssaho.2025.101721.

[69] J. Abdulbaqi and G. O. Uzun, "Effect of the Students' Test Anxiety Prevention Program on Students' Experiences of Test Anxiety: A Mixed-Methods Experiment," *Sage Open*, vol. 14, no. 4, Oct. 2024, doi: 10.1177/21582440241296609.

[70] W. Chaiyasit, K. Chomsuwan, and S. Chanchalor, "Hybrid Teaching Using Problem-Based Learning to Promote Self-Directed Learning Abilities of Students during the COVID-19 Pandemic," *Int. J. Learn. Teach. Educ. Res.*, vol. 22, no. 8, pp. 1–22, Aug. 2023, doi: 10.26803/ijlter.22.8.1.

[71] Y. Zhao, S. Mohd Yusof, and M. Hou, "Developing through Merits and Demerits: A Literature Review of Online Classroom, Teaching Motivation and Teaching Methods in China Primary School," *Int. J. Learn. Teach. Educ. Res.*, vol. 22, no. 8, pp. 183–202, Aug. 2023, doi: 10.26803/ijlter.22.8.11.

[72] L. Darling-Hammond, L. Flook, C. Cook-Harvey, B. Barron, and D. Osher, "Implications for educational practice of the science of learning and development," *Appl. Dev. Sci.*, vol. 24, no. 2, pp. 97–140, Apr. 2020, doi: 10.1080/10888691.2018.1537791.

[73] G. K. Tsismalidou and B. K. Kondilis, "The Effects of Stress and Fatigue on Levels of Anxiety in Pilots: An Aviation Industry Sample," *J. Organ. Psychol.*, vol. 24, no. 2, Jul. 2024, doi: 10.33423/jop.v24i2.7084.