Journal of Public Transportation Community
Vol. 6 No. 2, October 2026, pp. 131-145
Online: https://ejournal.poltekbangsby.ac.id/index.php/jptc/about
e-ISSN :2808-3393

Rater-Mediated Dependability in a Six-Dimension Aviation English Performance Assessment for Air Traffic Control Students

Laila Rochmawati[1], Harunur Rosyid[2], Delmio A. M. G. de Sousa[3], Tiara Sylvia [4]

[1] Politeknik Penerbangan Surabaya, Surabaya, Indonesia
[2] Informatics Engineering, Universitas Muhammadiyah Gresik, Gresik, East Java, Indonesia
[3] The Administration of Airports and Air Navigation of Timor-Leste (ANATL E.P.), Timor-Leste
[4]Politeknik Penerbangan Medan, Medan, Indonesia
Corresponding author: Laila Rochmawati, lailarochmawati@poltekbangsby.ac.id, Indonesia

Abstract

Rater-mediated scoring is central to performance assessment, yet high reliability alone does not show how dependable scores were produced. This secondary methodological analysis of aggregate audited quality-control outputs examines the quality-control architecture and documented inter-rater agreement of a locally developed Aviation English oral-aural performance assessment used with 34 fourth-semester Air Traffic Control students in Indonesian vocational higher education. The assessment represented six analytic dimensions: pronunciation, structure, vocabulary, fluency, comprehension, and interaction. Two raters coded R1 and R2 completed a six-hour calibration session on 25 June 2026 using 12 anchor recordings; the calibration ICC was .91 against a prespecified criterion of .80. Main scoring used blinded, randomized recordings, analytic descriptors, anchor-based drift checks, and an adjudication rule for material disagreement. The verified report documents 408 participant-time-dimension rating units and an overall ICC(2,1)=.982, 95% CI [.979, .985], with a mean absolute rater difference of 1.39 points on the 0-100 analysis scale. Results indicate very high pooled agreement but do not establish dimension-specific reliability, test-form equivalence, official ICAO proficiency levels, or operational licensing competence. Rater dependability in safety-sensitive language assessment is therefore best documented as an evidence chain linking construct definition, descriptors, calibration, blinded scoring, monitoring, and bounded interpretation rather than a single coefficient.

Keywords: Aviation English, inter-rater reliability, performance assessment, rater calibration, blind scoring

Article History:
Received: 3 July 2026 ; Reviewed : 14 July 2026; Accepted: 13 August 2026; Published : 1 October 2026
DOI: 10.46491/jptc.v6i2.2872

How to cite : Rochmawati, L., Rosyid, H., de Sousa, D. A. M. G., & Sylvia, T. (2026). Rater-Mediated Dependability in a Six-Dimension Aviation English Performance Assessment for Air Traffic Control Students. ***Journal of Public Transportation Community*****, 6(2), 131–145.** https://doi.org/10.46491/jptc.v6i2.2872

© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution Share Alike (CC BY-SA) license (**https://creativecommons.org/licenses/by-sa/4.0/**)

Introduction

Performance assessment in Aviation English sits at the intersection of language testing, professional communication, and safety-sensitive vocational

education. This intersection creates a measurement problem that is different from the scoring of objective language items. When a learner's oral-aural performance is judged through scenario-based tasks, the resulting score is not produced by the task alone. It is produced through an interaction among the target construct, task design, analytic descriptors, recorded performance, and rater judgment. Reliability therefore depends not only on whether two raters eventually produce similar numerical values, but also on whether the scoring process constrains avoidable sources of rater variation in a transparent and auditable way (Bachman & Palmer, 2010; Douglas, 2000; McNamara, 1996).

This issue is especially consequential in Aviation English because familiar terminology can create an unjustified impression of regulatory equivalence. ICAO language proficiency requirements address operational communication and include six language-proficiency dimensions that have shaped training and assessment practices internationally (International Civil Aviation Organization [ICAO], 2010, 2022). Local educational assessments may legitimately use related dimensions to structure instruction or research, but this does not make them authorized ICAO tests. Research on Aviation English has repeatedly emphasized that domain definition, assessment purpose, standard setting, and local implementation must remain aligned (Alderson, 2009; Knoch, 2014; Moder & Halleck, 2009; Treadaway & Read, 2024; Yin et al., 2024). A defensible classroom or research assessment must therefore state what its scores represent and, equally, what they do not represent.

Aviation English assessment also belongs to the wider air-transport education and safety system. Radiotelephony communication between pilots and air traffic controllers is a safety-relevant element of air transportation, and training and communication practices are shaped by ICAO guidance (Farris et al., 2008; ICAO, 2009, 2023; Y.-h. Kim, 2023). Validity concerns about Aviation English tests have also been voiced from the perspectives of test-takers and assessors (Assassi & Ghodbane, 2023; H. Kim, 2018). Dependable scoring of the oral-aural performance of Air Traffic Control students is therefore relevant to the preparation of human resources for the transportation sector, which is the setting served by this journal.

Rater-mediated assessment introduces another layer of uncertainty. In contrast with objectively keyed responses, performance scores depend on human interpretation of descriptors and evidence in a response. Raters may differ in severity, in the aspects of performance they attend to, in how they resolve borderline cases, or in how they apply a descriptor after repeated exposure to similar performances (Eckes, 2015; Wind & Peterson, 2018). An inter-rater coefficient is useful because it summarizes agreement, but the coefficient is not self-interpreting. The choice of reliability model must fit the measurement design, and the reported coefficient should be read together with confidence

intervals, scoring procedures, and complementary indicators of disagreement (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Empirical studies of spoken-language scoring suggest that rater training and experience are associated with rater performance, and that different types of rater experience relate differently to severity and consistency (Davis, 2016; Lumley & McNamara, 1995; Neittaanmäki & Lamprianou, 2024; Weigle, 1998).

For absolute-agreement purposes, an intraclass correlation coefficient based on a two-way random-effects, single-measure formulation is appropriate when raters are treated as a sample from a broader rater population and the numerical equality of scores matters, not merely their rank ordering (Koo & Li, 2016; McGraw & Wong, 1996). Yet even an excellent ICC cannot demonstrate that the rubric represents the intended construct. Two raters can agree closely while applying an underspecified or construct-irrelevant rubric. This distinction parallels contemporary validity arguments in which evidence about scoring and generalization is necessary but insufficient for broader interpretation and use claims (Bachman & Palmer, 2010; Flake et al., 2017; Kane, 2013; Messick, 1989).

The parent Aviation English project from which the present analysis is drawn developed a scenario-based educational performance assessment for Air Traffic Control students. The construct was defined as oral-aural Aviation English communication performance on research scenarios and was represented by pronunciation, structure, vocabulary, fluency, comprehension, and interaction. The construct was deliberately bounded as an educational research outcome rather than operational competence, licensing status, or official ICAO certification. Performance was recorded and evaluated by two raters under a structured quality-control protocol that included construct orientation, descriptor review, anchor scoring, independent calibration, blinded main scoring, drift checks, and adjudication rules.

The wider project generated several distinct evidence streams. Previous analyses from the project addressed comparative learning outcomes, dynamic-assessment process evidence, usability and implementation fidelity, instructional-design architecture, and a validity argument for an internal listening-reading assessment. The present study does not re-estimate those outcomes. Its unit of inquiry is the rater-mediated scoring process for the oral-aural performance assessment. This separation is important because reliability of human scoring is a measurement question in its own right and should not be treated as a minor statistic embedded inside an intervention paper.

Much of the Aviation English assessment literature addresses policy, test validity, and standard setting (Alderson, 2009; Assassi & Ghodbane, 2023; Knoch, 2014; Treadaway & Read, 2024; Yin et al., 2024). To the authors' knowledge, less attention has been given to documenting the scoring operation of locally developed educational assessments as a stepwise, auditable chain from construct

orientation to adjudication, in which an agreement coefficient is reported together with the controls that produced it. The gap addressed here is therefore not a new coefficient but an explicit account of how rater dependability was produced and which score interpretations that evidence can and cannot bear. The contribution is a documented quality-control architecture, paired with bounded pooled agreement evidence, that vocational programs can adapt.

Three research questions guided the study: (RQ1) How was the rater-mediated scoring process structured to reduce preventable variation in a six-dimension Aviation English performance assessment? (RQ2) What level of pooled inter-rater agreement was documented for the 408 participant-time-dimension rating units? (RQ3) What score interpretations are supported, and what interpretations remain unsupported, by the available rater-quality evidence?

Research Method

Design and evidence base

This study used a secondary methodological analysis of verified rater-quality evidence archived in the final audited report of a larger Aviation English educational research project. The purpose was not to reconstruct participant-level rater data or to reanalyse the intervention effect. Instead, the analysis isolated the documented scoring architecture, calibration evidence, main inter-rater reliability estimate, mean absolute difference, and inferential boundaries attached to those results. No participant-level rater values were reconstructed or reanalysed for this paper; all numerical claims were restricted to aggregate quality-control outputs verified in the final audited report.

The parent study was conducted in an aviation-vocational higher-education context with 34 fourth-semester students in an Air Traffic Control study program. All 34 participants completed the pretest and posttest and, according to the audited research report, provided informed consent and consent for voice recording. The present paper uses no participant names and refers only to the scoring design and aggregate quality-control results.

Assessment construct and rubric

The oral-aural construct was defined as the quality of performance in Aviation English research scenarios requiring message comprehension, spoken response, and management of interaction. Six analytic dimensions were used: pronunciation, structure, vocabulary, fluency, comprehension, and interaction. The dimensions were construct-referenced to Aviation English communication requirements, but they were not used to assign official ICAO proficiency levels. The report further states that a critical-information checklist supported

interpretation of communicative completeness but was not treated as a separate empirical outcome.

The analytic rubric used a five-level descriptor scale for each dimension. The descriptors emphasized intelligibility and functional communication rather than simple error counts. Pronunciation addressed intelligibility; structure addressed adequacy and flexibility of grammatical forms; vocabulary addressed range, precision, and reformulation; fluency addressed continuity and control; comprehension addressed accurate and timely understanding; and interaction addressed response management, clarification, repair, and mutual understanding. Anchor audio, decision rules, and a critical-information checklist were specified as supporting scoring materials.

Table 1. Analytic dimensions and bounded interpretation

Dimension	Primary scoring focus	Boundary
Pronunciation	Intelligibility of spoken production	Not native-likeness or accent elimination
Structure	Adequacy and flexibility of grammatical forms	Not isolated grammar-item accuracy
Vocabulary	Range, precision, appropriateness, reformulation	Not technical-knowledge recall
Fluency	Continuity, pacing, and control	Not speed alone
Comprehension	Accurate and timely understanding	Not a standalone listening test score
Interaction	Response management, clarification, repair, mutual understanding	Not operational licensing competence

Note. The six dimensions were used as educational research descriptors. They were not used to assign official ICAO proficiency levels.

Rater preparation and quality-control workflow

Two raters, coded R1 and R2, conducted the performance scoring. Both raters completed the same prespecified preparation and calibration sequence before main scoring and were required to meet the study calibration criterion. Rater codes are retained throughout this manuscript so that the analysis remains focused on the dependability of the scoring operation rather than on individual assessor identity. The audited report available for this analysis does not document the raters' qualifications or years of experience, and describes the anchor recordings only as representing the performance range; these details are therefore not reported here and are acknowledged as a limitation.

The quality-control sequence comprised seven stages. First, raters reviewed the construct, the six dimensions, and score-use boundaries. Second, they analysed descriptor levels and critical features. Third, they scored anchor

recordings representing the performance range. Fourth, they completed an independent, unlabeled calibration set. Fifth, main scoring used randomized recordings with participant identity, testing occasion, and study condition concealed from raters. Sixth, anchor material was inserted during scoring as a drift check. Seventh, cases exceeding the specified disagreement threshold were subject to an adjudication procedure with the reason for the decision recorded. The archived protocol defined a calibration requirement of at least ICC=.80 for the implemented calibration session and a broader main-study criterion of ICC(2,1) at least .75, with .90 as the target.

The implemented calibration session lasted six hours on 25 June 2026 and used 12 anchor recordings. The archived calibration ICC was .91, exceeding the stated .80 calibration threshold. This calibration result was treated as a gate before main scoring rather than as the final reliability estimate.

Table 2. Rater quality-control sequence

Stage	Documented activity	Decision or control
1. Construct orientation	Review domain, dimensions, and score-use boundary	Raters distinguish the six dimensions
2. Descriptor review	Analyse scale levels and critical features	Shared interpretation of descriptor language
3. Anchor scoring	Score representative anchor recordings	Calibration evidence before main scoring
4. Independent calibration	Score an unlabeled calibration set independently	Implemented gate: ICC at least .80
5. Blind main scoring	Randomized files; identity, occasion, and condition concealed	Reduce expectation and metadata bias
6. Drift check	Insert anchor material during scoring	Monitor systematic change in scoring
7. Adjudication	Review materially discrepant cases under a rule	Record reason for final decision

Note. The protocol specifies the workflow. The verified archive did not provide a quantitative drift-event count or adjudication frequency for this paper.

Reliability analysis

Main reliability was evaluated on 408 rating units defined by the cross-classification of 34 participants, two testing occasions, and six analytic dimensions. The final report specifies ICC(2,1), a two-way random-effects, absolute-agreement, single-measure coefficient. This formulation was selected because raters were conceptually treated as a sample from a broader rater population and absolute score agreement was required (McGraw & Wong, 1996). Mean absolute rater difference was reported as a complementary metric. The

final audited report labels the confidence interval as a bootstrap interval. The report available for this analysis does not specify the resampling replication count or interval algorithm; therefore, this paper reports the verified interval without adding unverified bootstrap settings.

The 408 units are crossed classifications of participants, occasions, and dimensions and are therefore not independent observations: scores from the same participant share a common performance level across dimensions and occasions. In ICC(2,1), each unit is treated as a target scored by the same two raters, so the coefficient summarizes agreement across the pooled set of targets. Pooling suits the question asked here, namely the overall dependability of the scoring operation, but the reported confidence interval should be read with this clustering in mind, and the pooled coefficient is not a substitute for dimension- or occasion-specific estimates or for a multilevel or generalizability model.

The same report presents Cronbach's alpha=.972 for the six-dimension rubric after averaging rater scores within each testing occasion. Because the six dimensions are analytically distinct components of performance, this coefficient is treated only as evidence of strong internal coherence in the observed score profile, not as proof that the construct is unidimensional. This interpretation follows cautions against treating coefficient alpha as a standalone construct-validity statistic (McNeish, 2018; Sijtsma, 2009).

The analysis for this paper was deliberately bounded. The verified study output selected for this paper consisted of the pooled ICC, its reported confidence interval, the mean absolute rater difference, the calibration evidence, and the documented quality-control workflow. Descriptor-specific ICCs, rater-severity estimates, generalizability coefficients, many-facet Rasch estimates, drift slopes, and adjudication frequencies were therefore not introduced as post hoc results. The paper distinguishes between procedures documented in the protocol and numerical results explicitly reported in the audited study outputs.

Results and Discussion

Results

The rater-quality architecture was designed as a sequence rather than a single end-point statistic. The scoring process began with construct and descriptor orientation, moved through anchor-based guided scoring and independent calibration, and then separated main scoring from study metadata through randomized blinded files. Drift monitoring and adjudication were specified as safeguards during the main scoring phase. This sequence created an audit path from construct definition to final scored data and made explicit where human judgment entered the assessment process.

The calibration gate was met. The two raters completed six hours of calibration on 25 June 2026 using 12 anchor recordings. The archived ICC for the calibration set was .91, above the prespecified .80 threshold. The calibration result therefore supported proceeding to main scoring under the documented protocol.

For main scoring, the verified report documents 408 participant-time-dimension rating units. The pooled absolute-agreement coefficient was ICC(2,1)=.982 with a reported 95% CI [.979, .985]. The mean absolute rater difference was 1.39 points on the 0-100 analysis scale used in the report. Under the interpretive bands used in the parent study, the ICC is in the excellent range. The narrow reported interval also indicates that the pooled agreement estimate was not highly unstable within the archived analysis.

Table 3. Verified rater-agreement evidence

Evidence	Verified value	Interpretation
Calibration duration and date	6 hours; 25 June 2026	Pre-main-scoring calibration
Anchor recordings	12	Used in implemented calibration
Calibration ICC	.91	Exceeded prespecified .80 gate
Main rating units	408 = 34 participants x 2 occasions x 6 dimensions	Pooled participant-time-dimension units
ICC model	ICC(2,1): two-way random, absolute agreement, single measure	Absolute-agreement reliability
Main ICC	.982	Very high pooled agreement
95% CI	[.979, .985]	Reported as bootstrap CI in audited report
Mean absolute difference	1.39 points	Average absolute score distance on reported 0-100 analysis scale
Rubric alpha	.972	High internal coherence; not proof of unidimensionality

Note. No dimension-specific ICC, drift frequency, or adjudication count is inferred from these aggregate results.

The six-dimension rubric showed Cronbach's alpha=.972 after rater scores were averaged within each testing occasion. This high coefficient indicates strong coherence among the observed dimension scores, but the project retained dimension-level reporting because pronunciation, structure, vocabulary, fluency, comprehension, and interaction are substantively distinct aspects of performance. The alpha result was therefore not used to collapse the construct into a single undifferentiated language factor.

The available evidence supports a bounded conclusion: the two-rater scoring operation produced highly consistent pooled scores across the set of

rated participant-time-dimension units represented in the archive. The evidence does not show that each dimension independently achieved ICC=.982, because no dimension-specific ICC was verified. It does not establish that pretest and posttest forms were psychometrically equivalent, and it does not authorize conversion of local scores into official ICAO proficiency levels.

Discussion

The central finding is not simply that the pooled ICC was high. The more important methodological point is that the coefficient sat within a layered scoring system. Calibration occurred before main scoring; main files were blinded and randomized; analytic descriptors separated dimensions of performance; anchor material was used for monitoring; and adjudication was governed by a predefined rule. This architecture matters because dependability in performance assessment is produced through a chain of controls. A strong coefficient observed after an undocumented scoring process is less informative than the same coefficient accompanied by evidence about how raters were trained, what they were allowed to see, how disagreements were monitored, and how score-use boundaries were defined.

The observed ICC(2,1)=.982 is consistent with very high pooled absolute agreement under commonly used interpretive guidelines (Cicchetti, 1994; Koo & Li, 2016). The mean absolute difference of 1.39 points complements the ICC by expressing disagreement in the score metric used for analysis. The two measures answer different questions: ICC summarizes the proportion of variance attributable to differences among rated units relative to disagreement and other sources of variation, whereas mean absolute difference communicates the typical numerical distance between rater scores. Reporting both prevents the reliability claim from depending entirely on a single scale-free coefficient.

The calibration result is also consequential. The .91 calibration ICC was obtained before main scoring and exceeded the predefined .80 gate. This makes calibration a decision point rather than a ceremonial training activity. In rater-mediated language assessment, training is most defensible when it demonstrates that raters can operationalize the rubric before their scores contribute to an analysis. The approach is compatible with the broader principle that reliability evidence should be planned as part of assessment design rather than calculated only after data collection (American Educational Research Association et al., 2014; Koo & Li, 2016; Shrout & Fleiss, 1979). Empirical work on spoken-language scoring suggests that training and experience are associated with rater performance and that types of experience relate differently to severity and consistency (Davis, 2016; Neittaanmäki & Lamprianou, 2024), which is consistent with treating calibration as an explicit gate with a stated criterion rather than as general preparation.

Blinding is particularly valuable in intervention research. If raters know that a recording belongs to an experimental group or to a posttest occasion, expectation can enter the scoring process even when raters intend to be objective. The parent protocol concealed participant identity, study condition, and testing occasion during main scoring. This does not eliminate all rater effects, but it removes avoidable contextual information that could create differential scoring. In that sense, blind scoring functions as a measurement control analogous to other bias-reduction procedures in experimental research (Hróbjartsson et al., 2012).

The six-dimension design also reflects the specialized nature of Aviation English. Aviation communication cannot be reduced to generic grammatical accuracy. Intelligibility, lexical precision, fluency, comprehension, and interactional repair can all matter in professional communication, and the salience of these components is shaped by domain tasks (Douglas, 2000; Estival et al., 2016; Farris et al., 2008; H. Kim, 2018; Kim & Elder, 2009; Munro & Derwing, 1995). The project therefore treated pronunciation, structure, vocabulary, fluency, comprehension, and interaction as analytically separable. This is compatible with ICAO's language-proficiency framework while remaining intentionally distinct from official licensing decisions (ICAO, 2010, 2022).

The high alpha of .972 requires caution rather than celebration alone. A very high internal-consistency coefficient may reflect a strong common performance factor, but it can also arise because dimensions are closely related or partly redundant. Alpha does not establish dimensionality, and removing dimensions merely to maximize internal consistency can reduce construct coverage (McNeish, 2018). The parent project appropriately retained the six analytic dimensions. For future validation, factor-analytic or generalizability-theory evidence would be more informative for separating common performance variance from dimension-specific and rater-specific sources (Brennan, 2001).

The study also illustrates why rater reliability should not be confused with validity. Close agreement shows that raters applied the scoring system consistently, but it does not by itself show that the tasks sampled the full target domain, that the rubric captured every relevant feature of Aviation English, or that the resulting scores predict operational performance. Argument-based assessment practice requires a chain from domain definition and response evaluation to generalization, explanation, extrapolation, and use (Bachman & Palmer, 2010; Kane, 2013). The present evidence is strongest for the evaluation and rater-generalization portions of that chain. Broader claims require additional evidence.

This boundary is particularly important in Aviation English, where assessment results can be interpreted in relation to professional safety and regulation. The present scores were local educational research scores. They were

aligned with six Aviation English dimensions but were not official ICAO Language Proficiency Rating Scale decisions. Research on Aviation English test implementation and standard setting demonstrates that regulatory use requires much more than a reliable local rubric: test specifications, task representativeness, trained and governed rater systems, standard-setting evidence, security, consequences, and institutional authorization are all relevant (Alderson, 2009; Moder & Halleck, 2009; Treadaway & Read, 2024; Yin et al., 2024).

The findings have practical implications for vocational language programs that use locally developed speaking or radiotelephony tasks. First, rater training should begin with construct boundaries, not just score examples. Second, anchor performances should represent meaningful points on the scale and be accompanied by rationales. Third, independent calibration should be treated as a gate with a predefined criterion. Fourth, main scoring files should be stripped of metadata that can reveal condition or occasion where blinding is feasible. Fifth, monitoring should continue after calibration because agreement can drift during long scoring runs. Sixth, adjudication should leave an audit trail rather than silently overwrite original ratings. These controls are transferable even when the specific Aviation English rubric is not.

The study has several limitations. Only two raters were involved, and their qualifications and experience are not documented in the audited report, so the analysis cannot characterize a wider distribution of rater severity or rater-by-task interaction patterns. The pooled ICC combines 408 participant-time-dimension units that are not independent (see Reliability analysis) and should not be read as six separate dimension-specific reliability coefficients. The verified outputs also do not quantify drift-event frequency or adjudication counts, and the exact bootstrap resampling settings underlying the reported interval are not specified in the final analytical documentation used for this article. Accordingly, the present analysis does not fit a new rater-level model or infer unreported dimension-specific statistics. These limitations constrain decomposition of the reliability mechanism while leaving the reported pooled agreement evidence intact.

Future studies should preserve independent pre-adjudication ratings in a frozen, de-identified dataset and publish a reproducible analysis script. A stronger design would estimate reliability by dimension and testing occasion, examine systematic rater severity, and model participants, tasks, dimensions, occasions, and raters simultaneously. Generalizability theory or many-facet Rasch measurement could help distinguish task, rater, and dimension contributions to score variance (Brennan, 2001; Eckes, 2015). Preplanned drift analyses should report anchor performance over scoring batches, and adjudication logs should quantify both frequency and reason. Such extensions

would move the evidence from high pooled agreement toward a fuller measurement model.

A further priority is to connect rater dependability with consequences. In low-stakes classroom use, the principal concern may be whether scores support fair feedback and progress monitoring. In higher-stakes contexts, the evidentiary threshold is substantially higher because score differences can affect access, certification, or professional opportunities. The same numerical ICC may therefore be adequate for one use and insufficient for another. Dependability should be interpreted in relation to the decision attached to the score, not as a universal property of the rubric (Kane, 2013; Messick, 1989).

Conclusion

This study isolated the rater-mediated scoring evidence from a larger Aviation English educational research project and examined it as a measurement process in its own right. The verified evidence shows that two raters completed a six-hour, 12-anchor calibration session with ICC=.91 before main scoring and that 408 participant-time-dimension rating units yielded pooled ICC(2,1)=.982, 95% CI [.979, .985], with a mean absolute rater difference of 1.39 points. These values support very high pooled agreement for the implemented scoring operation. The stronger contribution, however, is procedural: construct orientation, analytic descriptors, calibration, blind randomized scoring, monitoring, and adjudication form a linked quality-control chain that makes rater dependability more auditable than a coefficient reported in isolation. The evidence must remain bounded. It does not establish dimension-specific reliability, form equivalence, official ICAO proficiency levels, or operational licensing competence. For Aviation English and other safety-sensitive forms of vocational ESP assessment, responsible score use requires both strong agreement evidence and explicit limits on what that evidence can justify.

Acknowledgements

The author expresses gratitude to the agencies, companies, and institutions that have offered support in conducting research and/or writing publications

AI Declaration

The author acknowledges the use of artificial intelligence (AI) tools in supporting the preparation of this manuscript. The author remains fully responsible for the accuracy, originality, and integrity of the research and its findings.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received for this work.

References

Alderson, J. C. (2009). Air safety, language assessment policy, and policy implementation: The case of aviation English. *Annual Review of Applied Linguistics, 29*, 168–187. https://doi.org/10.1017/S0267190509090138

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Assassi, T., & Ghodbane, T. (2023). The recurring issue of aviation English test validity: Echoes from test-takers and assessors of the English for Aviation Language Testing System in Algeria. *English Studies at NBU, 9*(2), 239–269. https://doi.org/10.33919/esnbu.23.2.6

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment,* 6(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.

Estival, D., Farris, C., & Molesworth, B. (2016). *Aviation English: A lingua franca for pilots and air traffic controllers*. Routledge.

Farris, C., Trofimovich, P., Segalowitz, N., & Gatbonton, E. (2008). Air traffic communication in a second language: Implications of cognitive factors for training and assessment. *TESOL Quarterly,* 42(3), 397–410. https://doi.org/10.1002/j.1545-7249.2008.tb00138.x

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social*

Psychological and Personality Science, 8(4), 370–378. https://doi.org/10.1177/1948550617693063

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*, 344, e1119. https://doi.org/10.1136/bmj.e1119

International Civil Aviation Organization. (2009). *Guidelines for Aviation English training programmes* (Circular 323-AN/185). ICAO.

International Civil Aviation Organization. (2010). *Manual on the implementation of ICAO language proficiency requirements* (Doc 9835, 2nd ed.). ICAO.

International Civil Aviation Organization. (2022). *Annex 1 to the Convention on International Civil Aviation: Personnel licensing* (14th ed.). ICAO.

International Civil Aviation Organization. (2023). *Annex 10 to the Convention on International Civil Aviation: Aeronautical telecommunications, Volume II—Communication procedures including those with PANS status*. ICAO.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement,* 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing,* 35(3), 403–426. https://doi.org/10.1177/0265532218758127

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics, 32*(3), 23.1–23.17. https://doi.org/10.2104/aral0923

Kim, Y.-h. (2023). The challenges of radiotelephony communication and effective training approaches: A study of Korean pilots and air traffic controllers. *English for Specific Purposes, 72*, 26–39. https://doi.org/10.1016/j.esp.2023.07.001

Knoch, U. (2014). Using subject specialists to validate an ESP rating scale: The case of the International Civil Aviation Organization (ICAO) rating scale. *English for Specific Purposes,* 33, 77–86. https://doi.org/10.1016/j.esp.2013.08.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing,* 12(1), 54–71. https://doi.org/10.1177/026553229501200104

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods,* 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Moder, C. L., & Halleck, G. B. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics,* 32(3), 25.1–25.16. https://doi.org/10.2104/aral0925

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning,* 45(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Neittaanmäki, R., & Lamprianou, I. (2024). All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency. *Language Testing, 41*(3), 606–626. https://doi.org/10.1177/02655322241239362

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika,* 74(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Treadaway, M., & Read, J. (2024). Setting standards for a diagnostic test of Aviation English for student pilots. *Language Testing, 41*(3), 557–583. https://doi.org/10.1177/02655322231224051

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing,* 15(2), 263–287. https://doi.org/10.1177/026553229801500205

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing,* 35(2), 161–192. https://doi.org/10.1177/0265532216686999

Yin, Y., Yu, G., & Huang, D. (2024). Implementing the ICAO Language Proficiency Requirements: The localization of Aviation English testing. *Applied Linguistics, 45*(4), 699–717. https://doi.org/10.1093/applin/amad057