Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2811–2819
https://doi.org/10.54941/ahfe1008505

Continuous Competency Assurance Through Human-AI Teaming: Reimagining Instructor Development

Debra Henneberry[1], Dimitrios Ziakkas[2], and Konstantinos Pechlivanis[3]

[1]School of Aviation and Transportation Technology, Purdue University, West Lafayette, IN 47907, USA
[2]School of Future Transport Engineering, Coventry University, Coventry, CV1 5FB, UK
[3]Technical University of Crete, Chania, 73100, Greece

ABSTRACT

Competency-based training relies upon consistent instructor judgement, yet instructor calibration remains one of the weakest elements of many training systems. This paper proposes a continuous competency assurance framework integrating instructor recruitment, onboarding, calibration, recurrent standardisation and AI-enabled grading analytics. Artificial intelligence identifies grading trends, inter-rater reliability and behavioural divergence, while instructors retain responsibility for assessment and feedback. The proposed methodology transforms instructor standardisation from periodic compliance activity into continuous organisational learning. The framework supports airline operations, Advanced Air Mobility and future AI-enabled training ecosystems.

Keywords: Instructor standardisation, Inter-rater reliability, Competency-based training, Human-AI teaming, Learning analytics, Rater calibration, Advanced air mobility

INTRODUCTION

Competency-based training and assessment (CBTA) and evidence-based training shift pilot training from the completion of prescribed manoeuvres to the demonstration of observable behaviours across a defined set of competencies (ICAO, 2020; IATA, 2014). This shift places the instructor's judgement at the centre of the training system. Every grade informs a decision about a trainee's progression, and aggregated grades inform the design of future training programmes. If instructors observe and grade the same performance differently, both decisions are compromised.

Research on instructor ratings in aviation shows that this risk is real. In a special issue devoted to training instructors to evaluate aircrew performance, Baker and Dismukes (2002) identified rater reliability as a central unresolved problem, and Brannick, Prince and Salas (2002) found that agreement between instructors was high for crew responses to scenario events but low for crew resource management items and scales. Studies of flight examiners have since shown substantial variance in how experienced pilots rate the same performance (Mavin, Roth and Dekker, 2013). Kahneman, Sibony

Received September 25, 2026; Revised October 25, 2026; Accepted November 3, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

and Sunstein (2021) describe such unwanted variability in professional judgement as noise, and argue that it is often larger and less visible than bias. In most training organisations, instructor standardisation still relies on periodic workshops and checks, so noise and drift between standardisation events remain largely undetected.

Artificial intelligence (AI) and learning analytics offer a means of making instructor judgement visible continuously rather than periodically (Siemens and Long, 2011). They also introduce risks: automation bias (Mosier and Skitka, 1996), erosion of professional autonomy, and the ethical challenges of monitoring professionals (Slade and Prinsloo, 2013). Our previous work on CBTA, emerging technologies and AI in training has argued that such tools must strengthen rather than replace instructor expertise (Ziakkas, Henneberry and Flores, 2023; Henneberry and Keller, 2024; Henneberry et al., 2026). This paper therefore proposes a Continuous Instructor Competency Assurance (CICA) framework and addresses three research questions (RQs): how instructor development can be structured as a continuous lifecycle (RQ1); how AI-enabled grading analytics can detect calibration problems while instructors retain responsibility for assessment (RQ2); and whether such analytics can reliably distinguish common rater effects in a proof-of-concept demonstration (RQ3).

BACKGROUND

Rater Effects and Inter-Rater Reliability

Measurement research identifies a consistent set of rater effects: severity or leniency, in which a rater systematically grades higher or lower than peers; halo, in which an overall impression dominates ratings of distinct dimensions; restriction of range; and inconsistency, in which a rater's grades vary unpredictably (Myford and Wolfe, 2003). Inter-rater reliability is commonly quantified using intraclass correlation coefficients (ICCs), whose form must match the rating design (Shrout and Fleiss, 1979); values between 0.75 and 0.90 are conventionally interpreted as good reliability (Koo and Li, 2016). Aviation behavioural marker systems such as NOTECHS were developed with explicit attention to rater training and reliability (Flin et al., 2003), yet reliability is typically evaluated in research projects rather than monitored during routine operations.

Rater Training and Instructor Development

Meta-analytic evidence shows that rater training improves rating accuracy, with frame-of-reference training, which aligns raters' mental models of performance standards, producing the most consistent benefits (Woehr and Huffcutt, 1994; Roch et al., 2012). Instructor expertise, like other forms of expertise, develops through deliberate practice with informative feedback (Ericsson, Krampe and Tesch-Römer, 1993). Competency-based education in aviation further emphasises that assessment is a professional judgement exercised within a shared understanding of performance (Kearns, Mavin and Hodge, 2016). In practice, however, instructors often receive feedback

on their own grading only during infrequent standardisation events, which limits the opportunity for deliberate practice.

AI in Assessment and Human-AI Teaming

AI applications in education have focused on profiling, prediction, assessment and adaptive systems, while the role of educators has received comparatively little attention (Zawacki-Richter et al., 2019). Large language models create further opportunities and challenges for assessment (Kasneci et al., 2023). Work on teacher-AI complementarity shows that analytics tools are most effective when co-designed with teachers and when they support, rather than supplant, professional decisions (Holstein, McLaren and Aleven, 2019). Two opposing failure modes must be managed: over-reliance on algorithmic outputs (Mosier and Skitka, 1996) and algorithm aversion after observed errors (Dietvorst, Simmons and Massey, 2015). Regulation reinforces the need for care. The EU Artificial Intelligence Act classifies AI systems intended to evaluate learning outcomes as high-risk (European Parliament and Council, 2024), which implies requirements for human oversight, transparency and data governance in any AI-enabled assessment system.

METHODOLOGY

The study combined framework development with a computational proof-of-concept. First, an integrative review, a method suited to combining diverse empirical and theoretical sources (Whittemore and Knafl, 2005), synthesised research on rater effects, rater training, instructor development, learning analytics and Human-AI Teaming. Second, the review findings were organised into a lifecycle framework using the logic of double-loop organisational learning, in which feedback revises not only individual behaviour but also the standards and assumptions that govern it (Argyris and Schön, 1978). Third, the analytics component was demonstrated using synthetic grading data generated from a specified rater-effects model.

The synthetic dataset represented 12 instructors grading 24 recorded training scenarios on nine competencies using a five-point scale over four quarterly calibration rounds, yielding 10,368 grades. Seven instructors were modelled as calibrated. Five were assigned known effects: leniency (I03), severity (I07), halo (I10), random inconsistency (I11) and progressive leniency drift (I05). The analytics pipeline computed ICC(2,1) for the instructor pool (Shrout and Fleiss, 1979) and four indicators for each instructor: severity relative to the peer consensus, consistency with the consensus, a halo index based on mean inter-competency correlation, and drift across quarters. Indicators were standardised using robust z-scores, and instructors were flagged when any absolute value reached 2.5. An unsupervised isolation forest model (Liu, Ting and Zhou, 2008) was applied to the same indicators as a comparative machine learning approach. Because the data were synthetic, the demonstration tests whether the analytics can recover known effects, not how often such effects occur in practice. Table 1 summarises the method.

Table 1: Research method of the study.

Element	Design Decision	Rationale
Research approach	Framework development with computational proof-of-concept	Combines conceptual contribution with a test of analytic feasibility
Evidence synthesis	Integrative review (Whittemore and Knafl, 2005)	Integrates measurement, training, analytics and teaming literatures
Framework logic	Instructor lifecycle structured by double-loop learning (Argyris and Schön, 1978)	Feedback revises both instructor behaviour and organisational standards
Demonstration data	Synthetic grades: 12 instructors, 24 scenarios, 9 competencies, 4 quarters (10,368 grades)	Known injected rater effects allow detection to be verified
Analytics	ICC(2,1); robust z-scores for severity, consistency, halo and drift; isolation forest	Combines interpretable statistics with unsupervised machine learning
Evaluation criterion	Recovery of five injected effects without false flags among seven calibrated instructors	Tests sensitivity and specificity of the analytics
Limitation control	No real trainee or instructor data used; results labelled as proof-of-concept	Avoids privacy risk and over-generalisation

THE CONTINUOUS INSTRUCTOR COMPETENCY ASSURANCE FRAMEWORK

The CICA framework (Figure 1) arranges instructor development as a five-stage lifecycle connected by a continuous analytics loop. Stage 1, recruitment and selection, defines an instructor competency profile that includes observation, assessment and facilitation skills, and uses structured selection activities such as grading a recorded performance. Stage 2, onboarding and initial calibration, applies frame-of-reference training with recorded reference performances whose grades have been agreed by a standards panel (Roch et al., 2012), establishing each instructor's baseline.

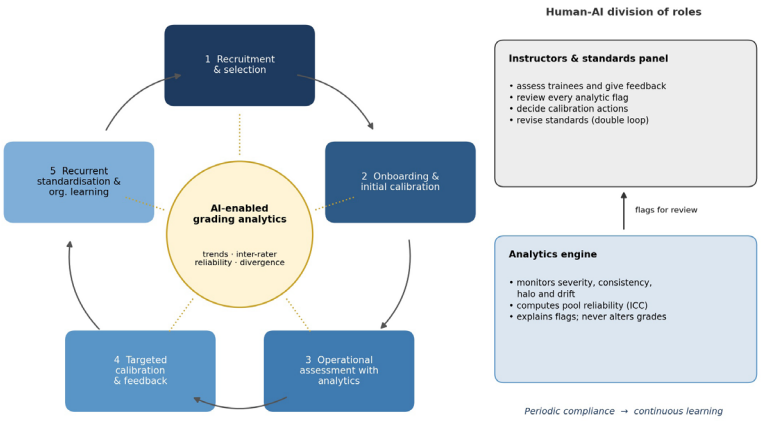


Figure 1: The continuous instructor competency assurance (CICA) framework.

Stage 3, operational assessment with continuous analytics, is the core innovation. As instructors grade trainees in routine training and periodically grade shared reference performances, the analytics engine monitors grading trends, inter-rater reliability and behavioural divergence. The engine does not grade trainees and does not alter grades; it produces flags and explanatory summaries for review by a standards instructor. Stage 4, targeted calibration and feedback, replaces uniform workshops with differentiated support: a lenient or severe instructor receives frame-of-reference recalibration, an instructor with halo effects receives training in behaviour-anchored observation, and an inconsistent instructor receives coaching and a review of workload during observation. Instructors see their own indicators, which turns analytics into deliberate practice feedback (Ericsson, Krampe and Tesch-Römer, 1993). Stage 5, recurrent standardisation and organisational learning, uses pool-level patterns to revise behavioural descriptors, scenario design and grading guidance, closing the double loop (Argyris and Schön, 1978).

Four principles govern the framework. Instructors retain responsibility for assessment and feedback, and every analytic flag is reviewed by a human before any action. Analytics are developmental rather than punitive, which protects the psychological safety needed for honest grading. Indicators are transparent and explainable to the instructors they describe. Data governance, including consent, access control and retention limits, meets the expectations for high-risk AI in education (European Parliament and Council, 2024).

PROOF-OF-CONCEPT FINDINGS

The analytics recovered all five injected rater effects and flagged none of the seven calibrated instructors, whose absolute robust z-scores did not exceed 1.3 on any indicator (Table 2). Severity and leniency were detected with large margins. Progressive drift in instructor I05 was detected through the drift indicator, with a slope of 0.23 grade points per quarter. Halo in I10 was detected primarily through reduced consistency with the consensus, while the halo index alone ($z = 1.9$) did not reach the threshold, suggesting that halo detection requires multiple indicators. The isolation forest flagged I05, I07, I10 and I11 but missed the lenient instructor I03, illustrating that an unsupervised model alone may overlook effects that interpretable indicators reveal clearly.

At pool level, ICC(2,1) was 0.69 in the first quarter and declined to 0.66 by the fourth quarter as drift accumulated. For the seven calibrated instructors alone, ICC(2,1) was 0.79, within the range conventionally described as good (Koo and Li, 2016). The difference illustrates how a small number of divergent instructors can reduce the reliability of a whole assessment system, and how targeted calibration could restore it without retraining the entire instructor pool.

Table 2: Findings: detection of injected rater effects in the synthetic proof-of-concept dataset.

Instructor and Injected Effect	Principal Indicator Value	Robust z-Score	Detected (Rule / Isolation Forest)	Framework Response
I03 leniency	Severity +0.61	19.7	Yes / No	Frame-of-reference recalibration
I07 severity	Severity −0.65	−18.7	Yes / Yes	Frame-of-reference recalibration
I05 progressive drift	Drift +0.23 per quarter; severity +0.39	10.0 (drift)	Yes / Yes	Early intervention between standardisation events
I10 halo	Consistency 0.78; halo index 0.71	−6.6 (consistency); 1.9 (halo)	Yes / Yes	Behaviour-anchored observation training
I11 inconsistency	Consistency 0.68	−13.9	Yes / Yes	Coaching; review of observation workload
Seven calibrated instructors	All indicators within normal range	$\lvert z \rvert \leq 1.3$	No / No	Continue routine monitoring

DISCUSSION

In response to RQ1, the CICA framework structures instructor development as a continuous lifecycle from recruitment to recurrent standardisation, linked by analytics and double-loop learning. In response to RQ2, the framework assigns AI the role of detecting grading trends, reliability problems and divergence, while instructors and standards personnel retain responsibility for assessment, feedback and any resulting action. In response to RQ3, the proof-of-concept showed that interpretable indicators can recover common rater effects without false flags, and that combining indicators is necessary for subtler effects such as halo.

The framework extends aviation research on instructor ratings (Baker and Dismukes, 2002; Brannick, Prince and Salas, 2002) by moving reliability assessment from occasional studies into routine operation. It also reframes noise in professional judgement (Kahneman, Sibony and Sunstein, 2021) as an organisational learning signal rather than an individual failing. From a Human-AI Teaming perspective, the design responds to both automation bias and algorithm aversion by keeping the analytics interpretable, reviewed and developmental (Mosier and Skitka, 1996; Dietvorst, Simmons and Massey, 2015). For airline operations, the framework offers a means of demonstrating the consistency of competency assessment to regulators and trainees. For Advanced Air Mobility, where instructor cadres will be new and small and

operational experience limited (Ziakkas, Henneberry and Pechlivanis, 2024), continuous calibration could establish consistent standards from the outset rather than after problems emerge. As training ecosystems adopt further AI tools, the same infrastructure could evaluate whether AI-generated assessment support is itself consistent with expert judgement.

The study has limitations. The demonstration used synthetic data with clearly separated effects, whereas real instructors may combine effects, and real reference performances may be graded against uncertain standards. The halo index is sensitive to genuine correlations between competencies. The framework has not been evaluated with instructors, whose acceptance will depend on trust and organisational culture. Future research should apply the analytics to anonymised operational grading data, evaluate instructor acceptance through participatory design, and test many-facet measurement models that estimate rater, trainee and competency effects simultaneously (Myford and Wolfe, 2003).

CONCLUSION

Competency-based training is only as reliable as the judgement of the instructors who assess it. The Continuous Instructor Competency Assurance framework integrates recruitment, onboarding, calibration, recurrent standardisation and AI-enabled grading analytics so that instructor consistency is monitored continuously and supported developmentally. The proof-of-concept indicates that the analytics can distinguish common rater effects, while the framework ensures that instructors retain responsibility for assessment and feedback. In doing so, instructor standardisation becomes continuous organisational learning that can serve airline operations, Advanced Air Mobility and future AI-enabled training ecosystems.

ACKNOWLEDGMENT

The authors thank colleagues at Purdue University, Coventry University and the Technical University of Crete for discussions that informed this work. No human participant data were used in this study.

REFERENCES

Argyris, C. and Schön, D.A. (1978). Organizational Learning: A Theory of Action Perspective. Reading, MA: Addison-Wesley.

Baker, D.P. and Dismukes, R.K. (2002). A framework for understanding crew performance assessment issues. International Journal of Aviation Psychology, 12(3), pp. 205–222.

Brannick, M.T., Prince, C. and Salas, E. (2002). The reliability of instructor evaluations of crew performance: Good news and not so good news. International Journal of Aviation Psychology, 12(3), pp. 241–261.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), pp. 114–126.

Ericsson, K.A., Krampe, R.T. and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), pp. 363–406.

European Parliament and Council (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L series, 12 July 2024.

Flin, R., Martin, L., Goeters, K.-M., Hörmann, H.-J., Amalberti, R., Valot, C. and Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. Human Factors and Aerospace Safety, 3(2), pp. 97–119.

Henneberry, D. and Keller, J. (2024). AI-CBTA impact on academia. In: Ziakkas, D., Plioutsias, A. and Pechlivanis, K. (eds.) Aviation Workforce Planning: The Artificial Intelligence and Competency-Based Era. HF Horizons, pp. 55–65.

Henneberry, D., Ziakkas, D., Lekea, I. and Pechlivanis, K. (2026). Harnessing emerging technologies to enhance decision-making in competency-based training and assessment. In: Intelligent Human Systems Integration (IHSI 2026): Disruptive and Innovative Technologies. AHFE International, 200.

Holstein, K., McLaren, B.M. and Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Journal of Learning Analytics, 6(2), pp. 27–52.

IATA (2014). Data Report for Evidence-Based Training. Montréal: International Air Transport Association.

ICAO (2020). Procedures for Air Navigation Services – Training (PANS-TRG, Doc 9868), 3rd edn. Montréal: International Civil Aviation Organization.

Kahneman, D., Sibony, O. and Sunstein, C.R. (2021). Noise: A Flaw in Human Judgment. London: William Collins.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Kearns, S.K., Mavin, T.J. and Hodge, S. (2016). Competency-Based Education in Aviation: Exploring Alternate Training Pathways. Farnham: Ashgate.

Koo, T.K. and Li, M.Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), pp. 155–163.

Liu, F.T., Ting, K.M. and Zhou, Z.-H. (2008). Isolation forest. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. Pisa: IEEE, pp. 413–422.

Mavin, T.J., Roth, W.-M. and Dekker, S. (2013). Understanding variance in pilot performance ratings: Two studies of flight examiners, captains, and first officers assessing the performance of peers. Aviation Psychology and Applied Human Factors, 3(2), pp. 53–62.

Mosier, K.L. and Skitka, L.J. (1996). Human decision makers and automated decision aids: Made for each other? In: Parasuraman, R. and Mouloua, M. (eds.) Automation and Human Performance: Theory and Applications. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 201–220.

Myford, C.M. and Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. Journal of Applied Measurement, 4(4), pp. 386–422.

Roch, S.G., Woehr, D.J., Mishra, V. and Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. Journal of Occupational and Organizational Psychology, 85(2), pp. 370–395.

Shrout, P.E. and Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), pp. 420–428.

Siemens, G. and Long, P. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review, 46(5), pp. 30–40.

Slade, S. and Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10), pp. 1510–1529.

Whittemore, R. and Knafl, K. (2005). The integrative review: Updated methodology. Journal of Advanced Nursing, 52(5), pp. 546–553.

Woehr, D.J. and Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology, 67(3), pp. 189–205.

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16, 39.

Ziakkas, D., Henneberry, D. and Flores, A. (2023). Competency-based training and assessment (CBTA) framework in global aviation training. Paper presented at the 14th International Conference on Applied Human Factors and Ergonomics (AHFE 2023), San Francisco, CA, 20–24 July.

Ziakkas, D., Henneberry, D. and Pechlivanis, K. (2024). The application of emerging technologies in electric vertical take-off and landing CBTA training. In: Human Factors in Design, Engineering, and Computing. AHFE International. https://doi.org/10.54941/ahfe1005776