

FROM DATA TO DEBRIEF: A FRAMEWORK FOR PILOT PERFORMANCE ASSESSMENT

Pedro Burth Kurka[1] & Emilia Villani[1]

[1]Aeronautics Institute of Technology (ITA), São José dos Campos, Brazil

Abstract

This paper proposes and demonstrates a framework for automated pilot performance assessment in flight simulators, grounded in Evidence-Based Training (EBT) principles. The system integrates flight data extraction via UDP protocol, automatic flight phase segmentation, and quantitative performance metrics aggregated into a phase-weighted composite score. The framework is applied to simulated traffic pattern flights with a T-27M Tucano (EMB 312) aircraft at the Brazilian Air Force Academy (AFA). Results demonstrate the internal consistency of the system, which correctly identified all flight phases and produced differentiated scores reflecting piloting differences across flights. The framework constitutes an instructional support tool that reduces subjectivity in traditional assessment, increases reproducibility, and delivers granular phase-level feedback.

Keywords: Evidence-Based Training; pilot assessment; flight simulator; performance metrics; composite score

1. Introduction

1.1 Background and Motivation

Aviation has come a long way in terms of safety in the last decades. However, despite all advances in reliability and automation, human factors remain a critical element and one of the most contributing sources for fatal accidents, with a large part of that attributable to non-technical skill deficiencies [1]. Flight training institutions traditionally follow teaching models focused primarily on lessons learned from accidents and incidents involving older aircraft, an approach no longer suited to modern operational risks. Against this backdrop, the International Civil Aviation Organization (ICAO) developed Evidence-Based Training (EBT), a competency-oriented methodology that uses operational data to identify, develop, and assess core pilot competencies [2].

1.2 The Assessment Challenge

Despite the conceptual advances of EBT, its implementation faces significant practical challenges [3, 4]. Flight instructors must be capable of assessing students across multiple competencies simultaneously such as observing crew resource management actions, inter-crew communication, instrument monitoring under complex conditions, in addition to quantitative parameters including airspeed, altitude, trajectory, and aircraft configuration [3]. Such assessment is conducted through limited samples of possible occurrences, which inevitably generates variability and potential evaluation errors.
Research shows that humans tend toward systematic biases in assessment, including serial position effects, the halo effect, and recognition-based inferences. Studies in maritime simulation have shown that conventional assessment methods present absolute reliability with coefficients of variation of up

DOI: https://doi.org/10.7195/icas2026_0692

to 50% for technical competencies. Consequently, these problems directly affect training effectiveness, as subsequent phases critically depend on the accurate identification of individual strengths and weaknesses [3, 4].
Pilot performance is typically assessed through isolated metrics such as path deviation, parameter maintenance, or landing quality. However, constructing a holistic, competency-based pilot evaluation invariably leads to biased interpretations. This gives rise to the need for a quantitative assessment model that avoids such instructor interference, while broadly capturing the desired performance profile [5], thereby complementing pilot evaluation and optimizing pilot development.

1.3 Research Objectives

The present paper aims to develop a framework for automated pilot performance assessment based on flight simulator training. The framework objectives are:

- **Data integration**: extract and process flight simulator data encompassing temporal parameters, geographic coordinates, airspeed, altitude, aircraft attitudes, engine performance information, and control inputs;
- **Quantitative feedback**: implement performance assessment methods based on weighted error criteria and behavioral indicators [3];
- **Composite Scoring**: develop holistic metrics based on individual measures such as parameter maintenance, approach stability, and landing quality, presented in a report containing a unified score [5];
- **Automated Reporting**: generate, at the conclusion of each simulated flight, a report with graphical visual representations of performance.

2. Methodology

2.1 Framework Architecture

The proposed framework consists of four sequential stages designed to transform raw flight data into a performance assessment report as described below.
Stage 1: Data Extraction. Flight parameters are extracted via UDP protocol, capturing data and exporting it to a file that can be accessed and processed.
Stage 2: Temporal Data Processing. The parameters extracted from the simulated flight are organized and divided into seven flight phases: taxi, departure, climb, level-off (downwind), base leg, approach, and landing. During each stage, data is extracted according to its temporal relevance within the context and desired tasks.
Stage 3: Competency-Based Performance Mapping. At this stage, each flight parameter is analyzed using quantitative assessment algorithms (detailed in Section 2.2), to produce a performance score. Each assigned score is based on the competencies required for that flight phase in accordance with its complexity, as well as the behavioral indicators of aggressiveness and duty cycle.
Stage 4: Performance Report. At the conclusion of the simulated flight, the pilot receives a performance report containing graphical visual representations and performance scores.

2.2 Quantitative Assessment Model

Pilot performance is assessed through a series of statistical indices designed to compare the executed flight against pre-defined parameters from the operational manuals. For each measured parameter, specific analyses are conducted according to the following methodologies:

2.2.1 Parameter Maintenance Precision (Tracking Accuracy — T_f)

This index assesses the deviation between actual performance and the ideal flight profile [6] (altitude maintenance, airspeed, vertical speed).
Root Mean Square Error (RMSE) is the primary formula for quantifying continuous errors during a phase.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (TargetValue - ObservedValue)^2}{n}} \tag{1}$$

- Target Value: the ideal value (e.g. 1,000 ft in altitude).
- Observed Value: the actual value extracted from the simulator telemetry at instant i.
- n: the total number of data samples collected during the segment.

Due to the nature of the flight, it is necessary to observe different information for each phase to correctly evaluate it. During approach, for example, where the altitude is constantly changing, the best parameter to be observed is vertical speed. On the other hand, downwind phase, where the objective is to maintain fixed altitude and speed, those are the parameters that should be used for the analysis.
Then, the RMSE values are converted into normalized tracking sub-scores through a linear tolerance-based formulation and used as required by the phases. In segments where more than one parameter is analyzed, a weighted linear combination is applied.
RMSE for altitude, IAS and Vertical Speed (VS) are calculated by

$$T_{alt} = \max\left(0,\ 1 - \frac{RMSE_{alt}}{2 \cdot Alt_{tol}}\right) \tag{2}$$

$$T_{ias} = \max\left(0,\ 1 - \frac{RMSE_{ias}}{2 \cdot IAS_{tol}}\right) \tag{3}$$

and

$$T_{vs} = \max\left(0,\ 1 - \frac{RMSE_{vs}}{2 \cdot VS_{tol}}\right) \tag{4}$$

where $RMSE_{alt}$, $RMSE_{ias}$ and $RMSE_{vs}$ are the root mean square errors of altitude, indicated airspeed and vertical speed, respectively. Alt_{tol}, IAS_{tol}, and VS_{tol} are the corresponding operational tolerance limits.
Under this mapping, a sustained deviation equal to the tolerance limit yields a score of 0.5, marking the boundary of satisfactory performance, and the score reaches zero only at twice the tolerance, what reflects operational practice, in which flying at the tolerance limit is acceptable rather than a failure.
Then, for each phase, the tracking score is obtained through a weighted linear combination.

$$T_f = \omega_{alt} T_{alt} + \omega_{ias} T_{ias} + \omega_{vs} T_{vs}, \qquad \omega_{alt} + \omega_{ias} + \omega_{vs} = 1 \tag{5}$$

which ensures that the tracking score remains bounded in the interval $[0,1]$, with 1 corresponding to ideal performance and 0 corresponding to deviations greater than the prescribed tolerance.
The operational tolerance limits and weights for each phase are described in Section 3.
Departure Score: since departure is a transitional event rather than a continuous phase, its evaluation consists of the rotation speed, and the speed at which the landing gear is retracted (according to operational standards). Thus, the tracking score for departure consists of

$$T_{rot} = \max\left(0,\ 1 - k_{rot} \cdot \frac{\left|IAS_{rot} - IAS_{rotref}\right|}{2 \cdot T_{rot}}\right) \tag{6}$$

and,

$$T_{gear} = \begin{cases} 1, & 90 \leq IAS_{gear} \leq 100 \\ \max\left(0,\ 1 - 0.05\left(90 - IAS_{gear}\right)\right), & IAS_{gear} < 90 \\ \max\left(0,\ 1 - 0.10\left(IAS_{gear} - 100\right)\right), & IAS_{gear} > 100 \end{cases} \tag{7}$$

$$T_{fdep} = \omega_{rot} T_{rot} + \omega_{gear} T_{gear}, \qquad \omega_{rot} + \omega_{gear} = 1 \tag{8}$$

where IAS_{rot} and IAS_{gear} stands for the Indicated Airspeed (IAS) for rotation and the IAS when landing gear is retracted, respectively. IAS_{rotref} is the reference speed for rotation and k_{rot} is the penalization coefficient, with a value of 0.40. ω_{rot} and ω_{gear} are the defined weight for rotation and gear retraction, respectively.

Landing Score: for critical phases such as touchdown, the assessment may combine multiple instantaneous parameters, such as IAS and vertical speed at touchdown, as adapted from [7]:

$$Tf_{IAS} = \max\left(0,\ 1 - k_{ias} \cdot \frac{\left|IAS_{td} - IAS_{ref}\right|}{T_{ias}}\right) \tag{9}$$

where IAS_{td} is the indicated airspeed at touchdown, IAS_{ref} is the reference touchdown speed, T_{ias} is the operational tolerance limit (both defined by the operator), and k_{ias} is the penalization coefficient, with a value of 0.4 for this analysis.
The vertical speed component is defined as

$$Tf_{VS} = \max\left(0,\ 1 - k_{vs} \cdot \frac{\left||VS_{td}| - \left|VS_{ref}\right|\right|}{T_{vs}}\right) \tag{10}$$

where VS_{td} is the vertical speed at touchdown, VS_{ref} is the reference sink rate, T_{vs} is the operational tolerance limit (both defined by the operator), and k_{vs} is the penalization coefficient, with attributed value of 0.05.
This formulation ensures that airspeed remains the decisive factor while vertical velocity modulates the overall score. Scores range continuously from 0 to 1, with no abrupt collapses, reflecting the operational principle that marginal deviations do not render the landing categorically poor.
Thus, Landing Score integrates both components through a linear combination of the normalized values and their respective weights ω_{IAS} and ω_{VS}:

$$Tf_{land} = \omega_{IAS} Tf_{IAS} + \omega_{VS} Tf_{VS}, \qquad \omega_{IAS} + \omega_{VS} = 1 \tag{11}$$

2.2.2 Control Strategy (C_f)

These assess the physical and cognitive effort of the pilot in control inputs [8] (stick and rudder pedals).
Aggressiveness: measures the rate of change of control movements (how abrupt the pilot is):

$$Aggressiveness = \sqrt{\frac{1}{n-1}\sum_{i=2}^{n}\left(\frac{\delta_i - \delta_{i-1}}{t_i - t_{i-1}}\right)^2} \tag{12}$$

where, δ_i is the control position (stick/pedal deflection) at instant i, t_i is the time at instant i and n is the number of samples in the analysis.
Duty Cycle: measures the percentage of time in which the pilot is actively applying control inputs above a deadband (neutral noise). It uses the aggressiveness results, filtering them through a previously defined threshold and gives as output information on how much time the pilot is actuating on the commands:

$$Duty\ Cycle = 100\% \times \frac{1}{t_n - t_2}\sum_{i=2}^{n} x_i\,(t_i - t_{i-1}) \tag{13}$$

where

$$x_i = \begin{cases} 0 & \text{for } \dfrac{\delta_i - \delta_{i-1}}{t_i - t_{i-1}} < \text{noise threshold and } |\delta_i| < \delta_{\max} \\ 1 & \text{otherwise} \end{cases}$$

Power Spectral Density (PSD). The PSD of the combined pitch–roll control input signal is obtained from the discrete Fourier transform in each flight phase, as in (14). The PSD-based control score is then defined as

$$X(n) = \sum_{m=1}^{N} x(m) e^{-j\frac{2\pi}{N}(m-1)(n-1)} \tag{14}$$

$$PSD = \frac{|X(n)|^2}{N}. \tag{15}$$

In (14), $x(m)$ is the instantaneous RMS combined values of pitch–roll signals, N denotes the number of samples in the analysis window used to compute the discrete Fourier transform. In this work,

the Short Time Fourier Transform (STFT) is used in time windows of N samples. The spectrum is normalized by dividing $|X(n)|^2$ by N so that the integral of $PSD(f)$ over frequency is proportional to the average signal energy. This makes PSD-based scores comparable across flight phases of different duration and is consistent with standard FFT-based PSD estimation methods.
The PSD-based control score is then defined as

$$C_{PSD} = \frac{\int_0^{0.25\,\mathrm{Hz}} PSD(f)\,df}{\int_0^{f_{Nyq}} PSD(f)\,df} \tag{16}$$

where f_{Nyq} is the Nyquist frequency. Following Hebbar and Pashilkar [8], who relate low-frequency control energy to smooth, anticipatory piloting, a cutoff of 0.25 Hz was adopted as the boundary between low-frequency (smooth) and higher-frequency (reactive) control inputs.
These control indicators are then converted into normalized scores in the interval $[0,1]$ through linear tolerance-based equations, defined as

$$C_A = \max\left(0,\ 1 - \frac{A_f}{A_{tol}}\right) \tag{17}$$

where A_f is the phase aggressiveness and A_{tol} is the operational tolerance, and

$$C_{DC} = \max\left(0,\ 1 - \frac{\left|DC_f - DC_{ref}\right|}{DC_{tol}}\right) \tag{18}$$

where DC_f is the measured Duty Cycle, DC_{tol} is the tolerance and DC_{ref} is a nominal reference activity level (noise threshold).
Then, the score for control strategy (C_f) for each phase is obtained as the weighted linear combination of control smoothness, command activity and spectral behavior into a single normalized score

$$C_f = \omega_A C_A + \omega_{DC} C_{DC} + \omega_{PSD} C_{PSD}, \qquad \omega_A + \omega_{DC} + \omega_{PSD} = 1 \tag{19}$$

where ω_A, ω_{DC} and ω_{PSD} are the weights for each parameter.

2.2.3 Configuration Score (G_f)

Aircraft configuration compliance (landing gear and flap positions) is evaluated as a logical check against the specific phase reference states defined in MAPRO (2026) [9].
The configuration score is defined as

$$G_f = \begin{cases} 1, & \text{if landing gear and flap positions are both compliant throughout the phase} \\ 0.5, & \text{if only one of the two conditions is compliant} \\ 0, & \text{if neither condition is compliant} \end{cases} \tag{20}$$

2.2.4 Phase Composite Score

The present framework adapts the principle inspired by the model of Zabirov et al. [3], which proposes the assignment of weight coefficients to behavioral competency indicators to reduce the subjectivity of instructional assessment to an automated telemetry-based assessment context. While the original model operates on indicators observed by the instructor, the present model replaces human observation with objective, continuous metrics derived from flight data.
The phase composite score (S_p) is calculated through a weighted linear combination:

$$S_p = 100 \cdot (\omega_T \cdot T_f + \omega_C \cdot C_f + \omega_G \cdot G_f), \qquad \omega_T + \omega_C + \omega_G = 1 \tag{21}$$

where T_f represents the tracking score, C_f the control score, G_f the configuration score, and the weights of the components of S_p are arbitrarily established by the authors in the present work with the following values:

- $\omega_T = 0.6$;

- $\omega_C = 0.3$; and
- $\omega_G = 0.1$.

The total sortie score is then computed as the weighted sum of phase scores:

$$S_c = \sum_f w_f \cdot S_p, \qquad \sum_f w_f = 1 \tag{22}$$

where w_f reflects the operational criticality of the phase according to the adopted weights.

2.2.5 Functional Relationships

It is notable that the demands placed on the pilot vary drastically when analyzing the proposed flight profile according to the phase. It is highly desirable to maintain constant altitude and airspeed, for example, when the aircraft is in the level-off or downwind leg, allowing automated assessment through fixed reference values for RMSE calculation. However, in vertical-transition phases such as climb or approach, both altitude and airspeed vary naturally.

According to Smith & Caldwell [6], calculating altitude deviation during a climb constitutes a metric without practical meaning, as the parameter is constantly changing. To address this challenge in automated measurement, Fuller et al. [7] establish the concept of "Functional Relationships", in which assessment parameters and targets change or become dynamic according to the specific phase of flight, evaluating rates of change rather than static set-points.

Accordingly, the trajectory and configuration assessment replace static parameters with dynamic metrics when necessary. The values adopted for the present article, following the T-27M aircraft standards, taken both from the procedure's manual of the Brazilian Air Force Academy (MAPRO 2026) [9] and from an experienced instructor, are presented in Table 1.

2.2.6 Phase Segmentation — State Machine

Phase boundaries are determined by a finite-state machine with seven sequential states (taxi, departure, climb, downwind, base leg, approach and landing). State transitions are triggered by threshold conditions on airspeed, altitude, heading, climb rate, landing gear, and flap position, all derived from the MAPRO [9] operational parameters. Runway heading is automatically detected from the heading recorded during takeoff roll, and circuit direction (left or right) is inferred from the deviation between the climb heading and the runway heading. The climb rate signal is smoothed with a Gaussian kernel prior to state evaluation, to prevent false transitions caused by momentary noise.

3. Implementation Scenario

The methodology described above is applied to a single traffic pattern profile in the Brazilian Air Force Academy (AFA) scenario, with the T-27M Tucano aircraft in the Xplane-11 simulator. Ten flights were performed by a certified instructor of the aircraft, with deliberate variations in piloting and parameter maintenance to illustrate different assessment scores. The traffic pattern profile consists of a single departure, climb to traffic altitude (3,000 ft), downwind leg maintenance, base leg turn, final approach, and landing.

Flight parameters are extracted via UDP protocol, capturing data at a frequency of 10 Hz, and exporting it to a *.csv* file containing 23 parameters, as listed:

1. **Timestamp**
2. **Latitude**
3. **Longitude**
4. **Elevation**
5. **Pitch**
6. **Roll**
7. **Yaw**
8. **Indicated airspeed (IAS)**
9. **Baro altitude**
10. **G normal**
11. **G lateral**
12. **Slip angle**

13. **Yoke Pitch Ratio**
14. **Yoke Roll Ratio**
15. **Yoke Heading Ratio**
16. **Throttle Ratio**
17. **Flap Handle Ratio**
18. **Gear Position**
19. **Hsi Loc**
20. **Hsi Gs**
21. **Ng Percent (Engine Rotation)**
22. **Nh Percent (Prop. Rotation)**
23. **Torque**

Results from application of the state machine described in Section 2.2.6, to one of the specific flights, yield the resulting graphic flight path indicator shown in Figure 1. Values used for determining the threshold for each phase are shown in Table 1.

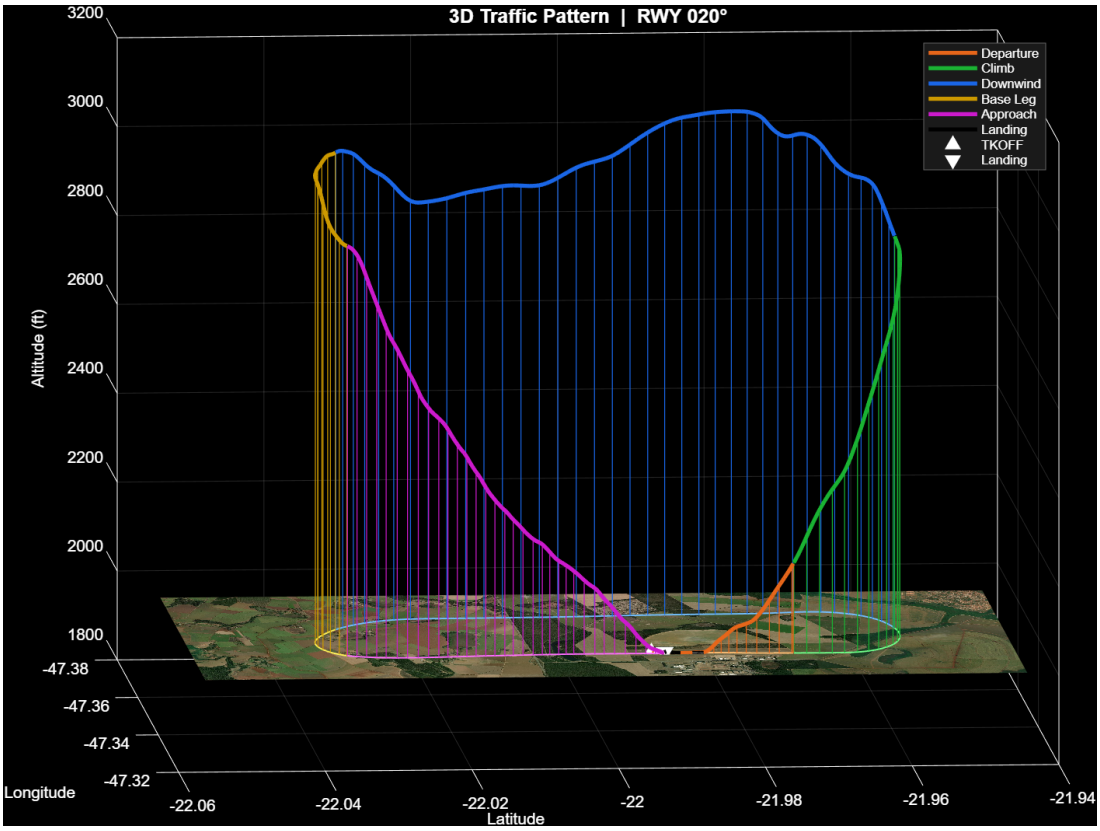


Figure 1 – Graphic traffic pattern indicator.

Table 1 – T-27M traffic pattern phase parameters (MAPRO 2026).

Phase	IAS (kt)	Climb	Ng (%)	Alt (ft)	Landing Gear	Flaps	Heading
Taxi	0–20	—	54 ± 1	1969 ± 10	Down	—	—
Departure	20–80	Positive	96 ± 3	> 1969	Transitioning up	TAKEOFF	Runway Centerline
Climb	80–140	Positive	96 ± 3	> 2100	Up	—	Runway heading / Turning
Downwind	180	—	—	3000 ± 200	Up	—	180° to runway
Base Leg	150–180	—	< 93	—	Transitioning down	—	Turning
Approach	< 150	Negative	—	—	Down	TAKEOFF – LANDING	Runway heading
Landing	< 90	Negative	—	< 2100	Down	– LANDING	Runway Centerline

The phase weights described in equation (22), used in the present evaluation examples, are defined by the author, based on his experience as a T-27M instructor and are shown in Table 2.
The reference values, tolerance limits, and weights adopted for each component of the assessment model are presented below. These parameters were established based on the operational standards defined in the MAPRO [9] and the author's experience as a T-27M flight instructor.

Table 2 – Phase weights (W_f) adopted for the present article.

Phase	$\mathbf{W_f}$
Taxi	0.00
Departure	0.15
Climb	0.15
Downwind / Level-Off	0.20
Base leg	0.10
Approach	0.25
Landing	0.15
Total	**1.00**

Table 3 – Tracking Score (T_f): phase-specific references and tolerances.

Phase	**Main Param.**	**Target**	**Tol.**	**Weight ω**	**Sec. Param.**	**Target**	**Tol.**	**Weight ω**
Departure	Rotation speed	80 kt	05 kt	0.50	Gear up speed	90 kt	10 kt	0.50
Climb	IAS	130 kt	25 kt	1.00	—	—	—	—
Downwind	Altitude	3000 ft	200 ft	0.70	IAS	180 kt	20 kt	0.30
Base Leg	Altitude	3000 ft	200 ft	1.00	—	—	—	0
Approach	IAS	110 kt	15 kt	1.00	—	—	—	—
Landing	IAS	90 kt	10 kt	0.75	Vertical Speed	-200 ft/min	150 ft/min	0.25

For each phase, parameters not listed carry an implicit weight of zero. The dominance of indicated airspeed (IAS) reflects its operational criticality during phases involving configuration changes (Departure, Climb, Approach), while altitude dominates in level segments (Downwind, Base Leg). The Base Leg is evaluated exclusively on altitude maintenance, as airspeed is deliberately reduced during this segment to configure the aircraft for landing prior to the glide path interception in the Approach phase.

For the Control Strategy (C_f) parameter, the stablished Aggressiveness tolerance ($A_{tol} = 1.0$) corresponds to the normalized aggressiveness magnitude established by Hebbar and Pashilkar [8]. The Duty Cycle reference ($DC_{ref} = 20\%$) represents an indicator of active piloting without excessive control effort, with a tolerance band ($DC_{tol} = 30\%$) that accommodates natural variations in pilot workload. Those values are represented on Table 4.

Table 4 – Control Score (C_f): component weights and tolerances.

Component	**Symbol**	**Reference**	**Tolerance**	**Weight**
Aggressiveness	C_A	—	$A_{tol} = 1$	$\omega_A = 0.30$
Duty Cycle	C_{DC}	$DC_{ref} = 20\%$	$DC_{tol} = 30\%$	$\omega_{DC} = 0.30$
PSD Score	C_{PSD}	—	—	$\omega_{PSD} = 0.40$

Choice of the PSD limit parameters, described in Section 2.2.2 were verified through the spectral analysis of different phases of the flight. Figure 2 shows the results of such an analysis.

4. Results and Discussion

4.1 Practical Implementation

The numerical analyses described in Section 3. are applied to the collected data from the simulated flights. The MATLAB computational platform is used to perform all calculations and provide the visual data associated with evaluation results.

The automated report generated by the framework is designed to be as simple and direct as possible, containing a summary of the flight and the distinct identified phases, followed by graphs representing the performance and scores assigned to the student according to the proposed methodology. Fig-

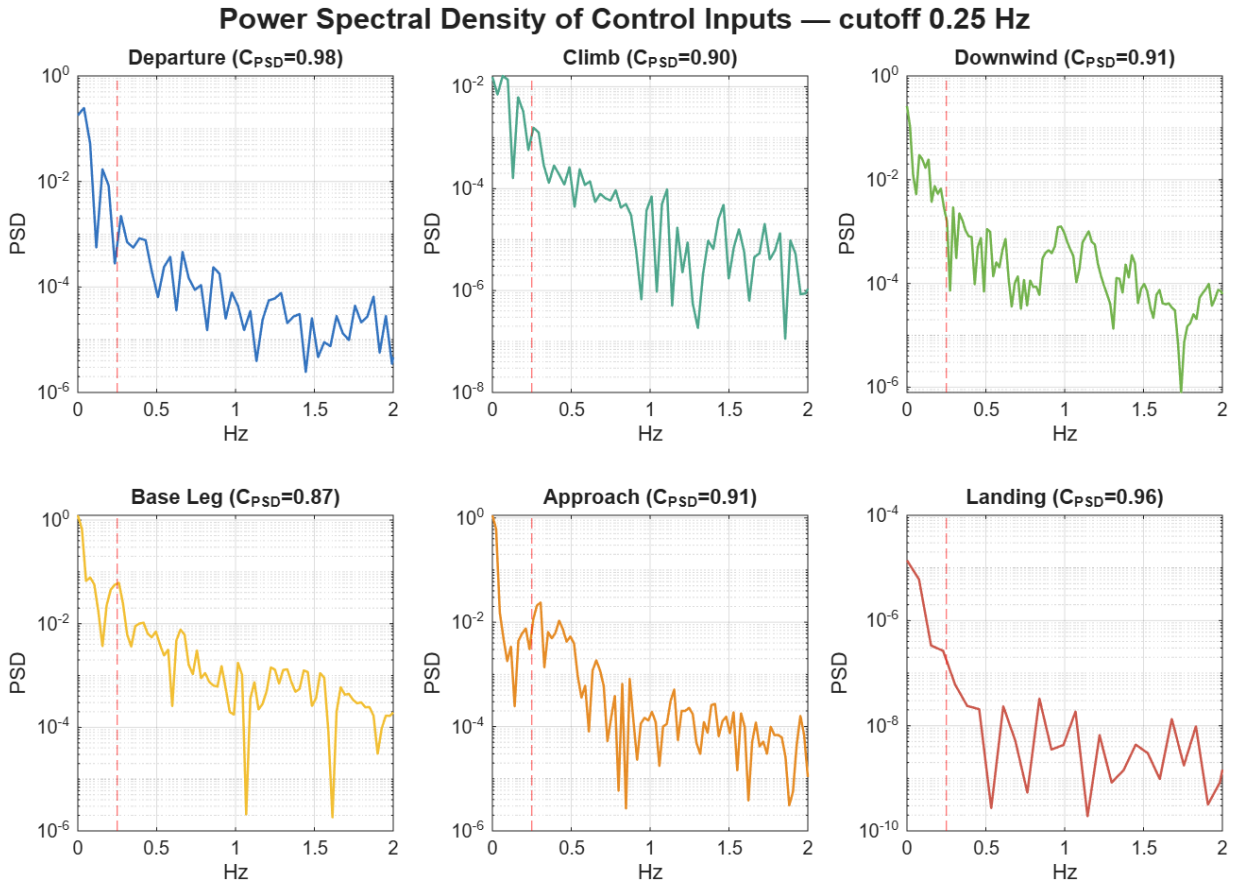


Figure 2 – Spectral analysis.

ure 3 shows the report dashboard providing the main classification and assessment information for the flight.

The initial flight analysis proved consistent with expectations: the framework correctly identified the different flight phases and applied the assessment methodology as intended.

During the execution of the assessment routine, certain difficulties were encountered, particularly in phase identification. Structuring heuristics robust enough to cover all cases and possibilities proves to be a challenge. However, it was possible to establish sufficient data to support the analysis by using the objective assessment data prescribed in the flight manuals, together with the knowledge and practical experience of specialists.

It is important to note that quantitative assessment presents a significant complexity due to the dynamic nature of flight. The parameters being evaluated are entirely different, depending on the phase, and the ultimate objective often goes beyond simple static parameter maintenance, encompassing the management of positioning, airspeed, configuration, climb or descent rate, and heading.

The analysis of simulated flights demonstrates that the proposed framework is capable of correctly segmenting flight phases and assigning differentiated scores according to the performance observed in each segment. The results reveal internal consistency among the metrics: flights with smaller altitude and airspeed deviations during the downwind leg and final approach produced significantly higher tracking scores (T_f), confirming the expected behaviour of the adopted function. Different flight evaluations can be seen in Figure 4:

The control dimension (C_f) demonstrates the capacity to distinguish piloting profiles with different degrees of aggressiveness and command effort. Flights with high-frequency and high-amplitude stick inputs produce lower control scores, consistent with the premise that efficient piloting is characterized by smooth and anticipatory corrections. The duty cycle parameter complements aggressiveness analysis by identifying the percentage of time with active control inputs by revealing patterns of pilot workload. Furthermore, the power spectral density (PSD) component adds a spectral dimension not captured by aggressiveness and duty cycle alone: it identifies whether control energy is concentrated

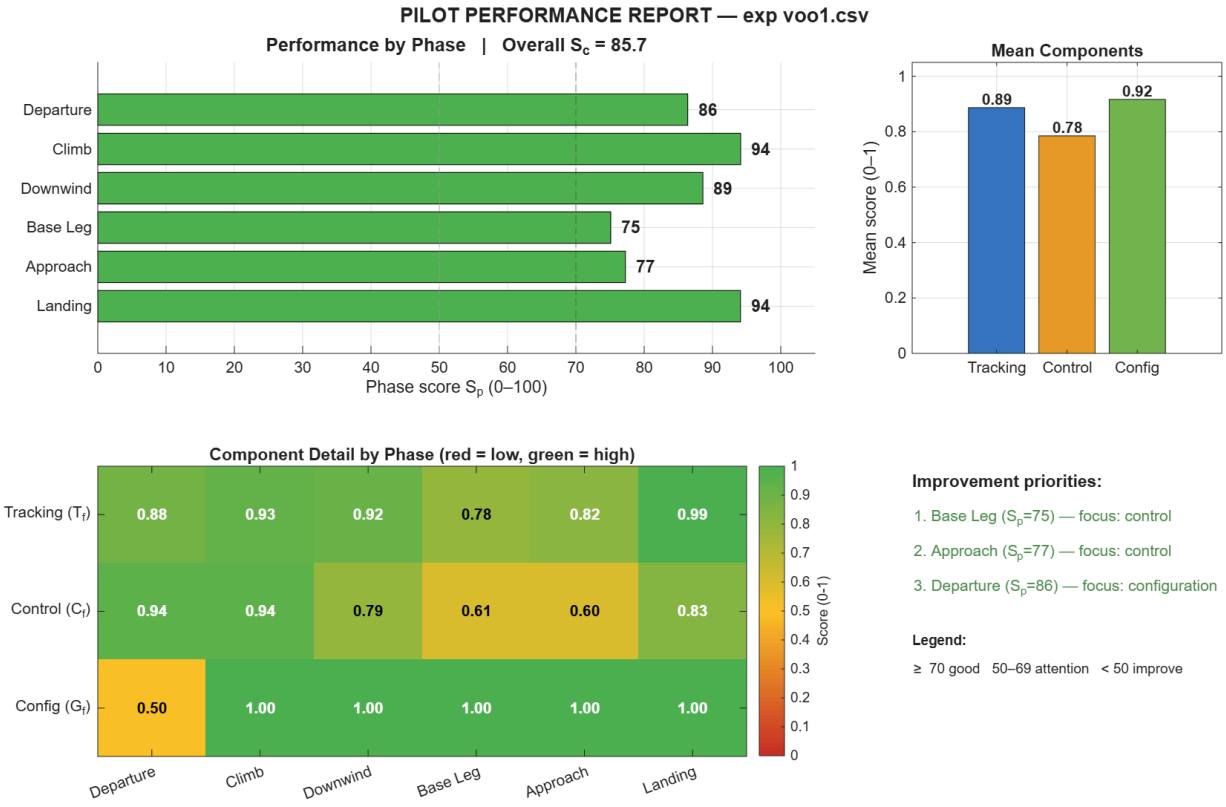


Figure 3 – Performance report.

at smooth, planned frequencies (below 0.25 Hz) or at reactive, high-frequency corrections. A PSD score close to 1.0 indicates that the pilot's energy is entirely in the low-frequency band, an indication of expert, anticipatory piloting.

The phase composite score structure (S_p), aggregated into a total score (S_c) through operational weights, proposes weight coefficients for behavioural indicators in EBT contexts. The primary distinction of the present work lies in replacing human observation with objective, continuous telemetry-based metrics, thereby eliminating inter-rater variability and the systematic biases identified in the literature, such as the halo effect and serial positioning effects [4].

The flights presented in this paper serve only as examples of the framework's application, and were all performed by the same pilot.

4.2 Comparative Evaluation with Instructor

To assess the validity of the framework, four experimental flights were independently evaluated by a qualified T-27 instructor pilot through a structured questionnaire. The instructor provided rating in a 0–10 scale evaluating each phase separately, together with an overall performance rating. The same four flights were then evaluated through the framework and the scores compared with the instructor ratings.

When comparing the assessments, although in different scales, a strong linear association was identified between both of them, with the evaluations classifying all the four flights in the same order, from the lower to the highest grades, indicating that the framework captures the same relative order of pilot performance that an experienced instructor perceives. The assigned grades are represented in Figure 5.

As shown, while flights were classified in the same order, the framework assigned much higher scores than the instructor, especially for the weaker flights, with the largest divergence occurring for the second flight, where the instructor attributed the lowest score (1 of 10), while the framework considered a score of 63.2 out of 100. This behavior was expected as the framework attributes grades following a linear basis, making a flight that is poor in absolute terms still accumulate partial scores across the analysis, which is consistent with the intended role of the framework to provide

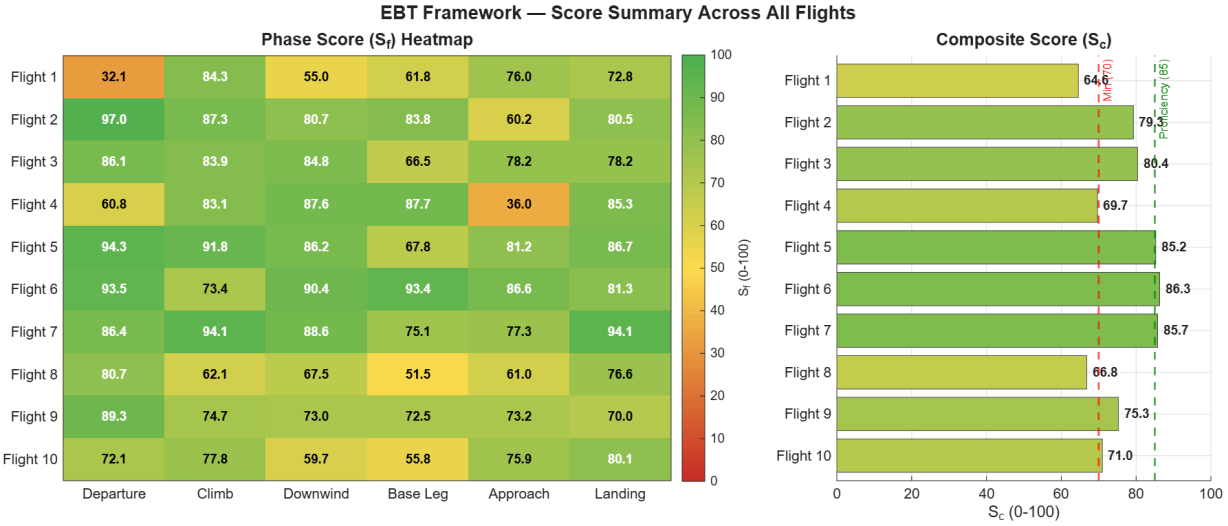


Figure 4 – Flight comparison.

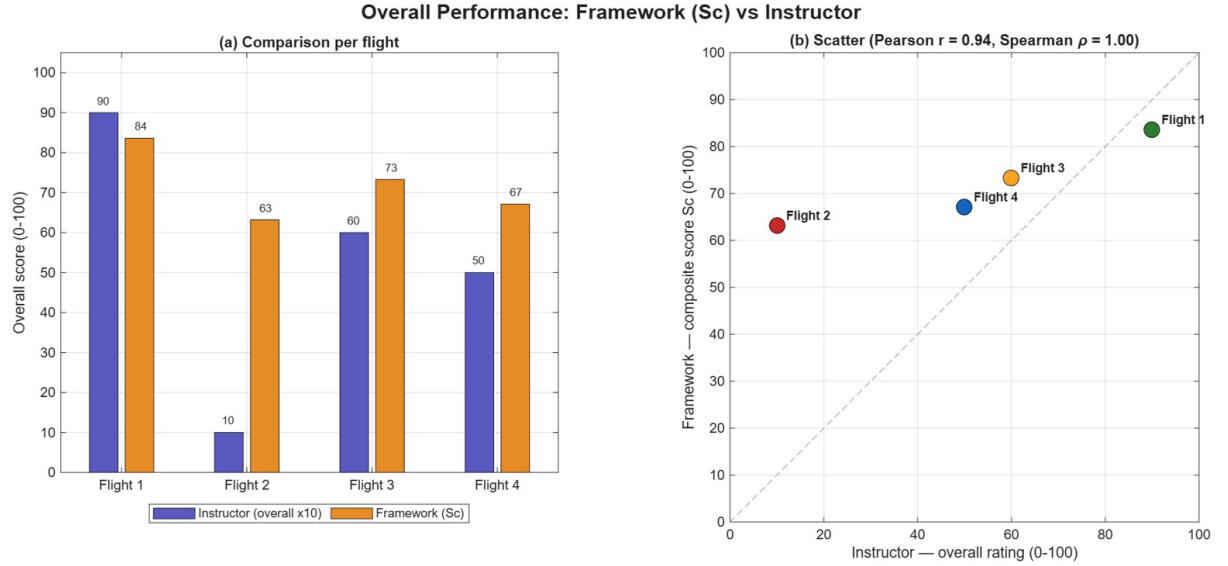


Figure 5 – Overall performance: framework (S_c) vs instructor.

support for debriefing, rather than a replacement for the instructors' judgement.
As for the Control Strategy, the framework precisely reflected the instructors' judgement, with a strong association between both assessments, as depicted in Figure 6, which represents the attributed grades of control strategy.

5. Considerations, Limitations and Next Steps

Implementing the framework in real operational contexts requires attention to a set of technical and institutional prerequisites. From a technical standpoint, result quality is directly dependent on telemetry data integrity and sampling rate. The 10 Hz UDP capture protocol used with the X-Plane 11 simulator proved sufficient for the evaluated metrics; however, for the analysis of flights with high-dynamic maneuvers, higher sampling rates are necessary to capture rapid variations in control inputs.
System parametrization requires operational domain expertise. The reference values for each phase (airspeeds, altitudes, landing gear and flap configurations), the tolerance limits used in the penalty functions, and the weights assigned to phases and dimensions must be defined by specialists familiar with the aircraft, the operational environment, and the specific instructional objectives. For this work, such parameters were established based on the author's experience as a T-27M instructor and the prescriptions of the MAPRO [9].

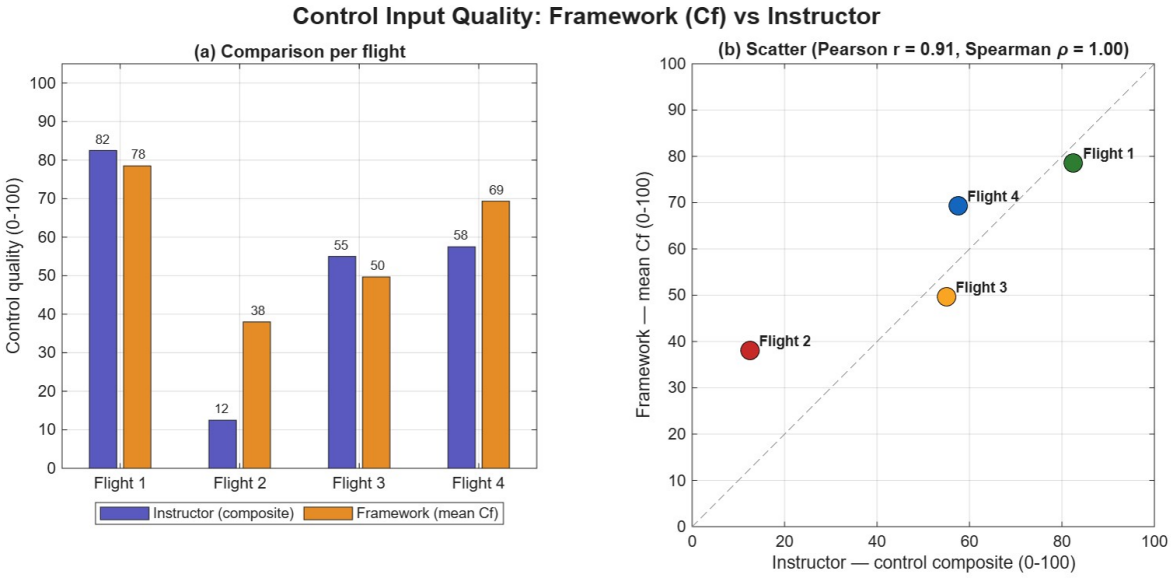


Figure 6 – Control input quality: framework (C_f) vs instructor.

From an institutional perspective, integrating the framework into formal training programs requires that automated scores be treated as instructor decision-support tools, not as a replacement for instructor judgement. The assessment of non-technical skills like crew resource management (CRM), decision-making under pressure, situational awareness, remains outside the scope of systems based exclusively on telemetry and demands qualified human observation.

Some limitations must be considered when interpreting results and extrapolating conclusions, such as:

Parameter calibration. Validation was conducted exclusively on a circuit landing traffic pattern profile, in a single geographic scenario (AFA) and with a single aircraft (T-27M). Generalization of the results to other mission types like aerobatics, cross-country navigation or formation flying requires the definition of new functional relationships, phase references, and specific tolerance limits, as discussed in Section 3.

Multi-rater and multi-pilot validation. The flights analyzed were performed by the author himself, and the methodology was not subjected to a wide external validation or to a representative sample of pilots at different proficiency levels. Future studies should conduct agreement analyses between framework scores and ratings by experienced instructors.

Phase identification robustness. The automatic phase segmentation via threshold-based heuristics proved functional for the analyzed profile, but its robustness was not tested under adverse conditions such as atypical piloting deviations, simulated equipment failures, or emergency maneuvers. Alternative segmentation methods may offer greater robustness and generality.

As perspectives for future research, the following are suggested:

1. The need for further studies and development on phases recognition;
2. Expansion of the framework as an automatic trainer, not only to evaluate the flight itself, but to be an adaptive learning module capable of automatically adjusting weights and thresholds based on the individual pilot's performance history to improve the abilities;
3. A study to evaluate how quantitative feedback interferes with the student progression along the flight training program.

6. Conclusion

The present paper proposed and demonstrated a framework for automated pilot performance assessment in flight simulators. The system integrates data extraction via UDP protocol, automatic flight phase segmentation, computation of quantitative metrics for tracking precision and control strategy,

and the generation of a phase-weighted composite score, synthesized in an automated report with visual performance representations.

The adopted methodology combines the use of Root Mean Square Error (RMSE) with linear tolerance-based normalization for tracking assessment, and the aggressiveness and duty cycle metrics for control strategy evaluation. The composite score structure adapts this principle to the context of continuous telemetry-based assessment, reducing dependence on human observation for the technical dimension of evaluation.

The results obtained with simulated flights in X-Plane 11, using the T-27M Tucano in the AFA traffic pattern, demonstrated the internal consistency of the system, producing scores that qualitatively reflect the piloting differences introduced across flights with correct phase identification and consistent application of the assessment methodology. The framework thus provides an instructional support tool that reduces the subjectivity inherent in traditional assessment, increases the reproducibility of the evaluation process.

The fact that the present work is limited to the analysis of a single, restricted flight profile, combined with the absence of validation by multiple specialists, delimits it as a methodological proof of concept, not a system validated for immediate operational use. Overcoming these limitations through the research perspectives outlined in the preceding section constitutes the development agenda for the framework to achieve sufficient maturity for institutional adoption in formal flight instruction programs.

In summary, the framework presented here demonstrates that automated quantitative assessment of pilots in simulators is technically feasible using the telemetry data already available in modern simulation systems, and that the appropriate methodological structuring of this information can contribute significantly to the effectiveness of evidence-based training.

7. Acknowledgments

The authors acknowledge the financial support of the following Brazilian Funding Agencies: FAPESP with grant 2021/11258-5; FINEP with grants 01.24.0600.00 and 01.20.0195.00. They also acknowledge the outstanding collaboration of Prof. Paulo Kurka (University of Campinas – UNICAMP) throughout the reviewing process of this paper.

8. Contact Author Email Address

Corresponding Author: kurkabr@gmail.com
Co-author: evillani@ita.br

9. Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] EASA. Terms of reference for a rulemaking task: Evidence-based and competency-based training. Technical report, European Aviation Safety Agency, 2018.

[2] ICAO. *Manual of evidence-based training*. International Civil Aviation Organization, 2013. Doc 9995.

[3] Z. Zabirov, A. Zabirov, and V. Shestakov. Modelling a quantitative assessment method of pilots' performance in evidence-based training. *Aviation*, 26(4):217–223, 2022. https://doi.org/10.3846/aviation.2022.18037.

[4] J. Ernstsen and S. Nazir. Performance assessment in full-scale simulators — a case of maritime pilotage operations. *Safety Science*, 129, 2020. https://doi.org/10.1016/j.ssci.2020.104775.

[5] Q. Chenot, F. Riedinger, F. Dehais, and S. Scannella. Assessing and visualizing pilot performance in traffic patterns: a composite score approach. *Safety*, 11(2), 2025. https://doi.org/10.3390/safety11020037.

[6] J. K. Smith and J. A. Caldwell. Methodology for evaluating the simulator flight performance of pilots. 2004.

[7] J. H. Fuller, W. L. Waag, and E. L. Martin. Advanced simulator for pilot training: design of automated performance measurement system. 1980.

[8] P. A. Hebbar and A. A. Pashilkar. Pilot performance evaluation of simulated flight approach and landing manoeuvres using quantitative assessment tools. *Sadhana*, 42(3):405–415, 2017. https://doi.org/10.1007/s12046-017-0613-0.

[9] MAPRO. *Manual de procedimentos operacionais — T-27M Tucano*. Academia da Força Aérea, 2026.