HANDLING QUALITIES RATING SCALE: THE IMPACT OF THE PILOT SUBJECTIVITY IN WORKLOAD EVALUATION

Mariane D. Turazza[1], João Paulo C. A. Macedo[2] & Jorge H. Bidinotto[3]

[1 2 3]São Carlos School of Engineering, University of São Paulo

Abstract

This work evaluates the impact of the pilot subjectivity in the workload evaluation using the Handling Qualities Rating Scale. Some emergency maneuvers were performed by flight test pilots in a moving base flight simulator, and the tasks were classified using the Cooper-Harper Scale. Results showed a strong variation in ratings for different pilots performing the same tasks, highlighting the importance of the development of a tool to become this evaluation free of the subjectivity caused by the pilots' opinion.

Keywords: Cooper-Harper Rating Scale, Pilot subjectivity, Human Factors, Flight Simulator

1. Introduction

During a flight test certification campaign of a new aircraft type, flying qualities evaluation is an underlying topic. As such, the pilots must be qualified to perform maneuvers during emergency situations. In other words, it is necessary to apply a qualitative analysis of the perception of pilots and not only quantitative with regard to the controllability of an aircraft, since the human aspect is extremely important especially when it comes to preserving the lives during a flight.

Workload measures the mental and physical efforts of pilots for attaining a given level of performance during a task. However, performance and workload are strongly connected, and cannot be characterised independently. To indirectly measure such parameters, Cooper, G. and Harper, R. proposed a method to evaluate this efforts, very widespread in aeronautical environment, called Handling Qualities Rating Scale [1].

This evaluation consist in a set of simple questions about the task accomplishment, that leads the result to a rating from 1 (with low workload) to 10 (high workload/loss of control). This number is known as Cooper-Harper Rating, or CHR. Thus, it becomes possible to have the average work effort to perform the maneuvers. In the last decades, the CHR became an accepted variable to accept or not the workload for an aircraft during its flight test certification campaign.

Variability in ratings also originates from the fact that each pilot is a unique individual, with varying abilities, decisions, and perceptions. In practice, this results in differences not only in the control strategies adopted by each pilot but also in the way they interpret the wording of the Cooper-Harper scale. Due to the peculiar nature of this subject, McDonnell, J. demonstrated that, from a psychological standpoint, the degree of compensation is a non-linear variable. Such differences in human perception can be reduced by selecting pilots with similar training, experience, and total flight hours.

Despite the practicality of working with numbers, an appropriate handling qualities assessment campaign involves much more than the analysis of CHRs. Numbers oversimplify a pilot's opinion because they do not explain how and why pilots decided on a rating; in fact, "only words can do that" [2].

Crew training in stall recognition and recovery, situational awareness, and the altitude at which maneuvers are performed helps explain the different outcomes and, therefore, should be part of a LOC-I definition. Noticing this gap, Bromfield, M. and Landry, S. conceived a qualitative re-definition of LOC-I based on "triggers" (factors potentially leading to a LOC-I) and "recovery factors" (aspects related to a possibly successful recovery). They concluded that a LOC-I "may be recoverable if recognized by the crew (situational awareness), given sufficient height above terrain and sufficient pitch, roll, and yaw control authority (controllability) for recovery within the airplane's structural design limits" [3].

This work aims to demonstrate, according to qualitative analysis, the pilots' perception regarding the workload applied during emergency situations, factors that demonstrate physical and/or mental exhaustion for each type of task to be performed. This is an extremely subjective scale, in which pilots must report how they feel in certain scenarios and maneuvers. Smaller discrepancies in workload analyzes are necessary, considering that we lose efficiency in terms of the real analysis in relation to the level of workload applied.

2. Methodology

In this work, some experiments were performed in a full motion flight simulator with 6 Degree of Freedom (Figure 1) using a Boeing 777 dynamics, inducing emergency situations in different scenarios, so that the pilot could recover the aircraft and be able to classify the degree of work effort for such maneuver, using the Cooper Harper scale. The tests were carried out with 6 experienced flight test pilots, capable of such classification. The same scenarios were exposed to all pilots, and they were induced to classify the task according to CHR, and the results were compared.

Figure 1 – Flight Simulator used for the tests.

After familiarizing the pilots with the operation of the aircraft, the emergency flight conditions were performed, which included engine fail, vehicle upset, ice conditions, windshear, unannounced failure in some systems, as flight controls, flight instruments, among other, in a total of 13 scenarios and 26 tasks, chosen among the most fatal and the most common, according to the statistics highlighted by [4]. In all of them, the pilots should recover the aircraft to a leveled and safe condition, and classify the maneuver, assigning a CHR level.

There are factors on a flight that contribute to analyse emergency conditions, for example, a night flight, turbulence or thunderstorm activity can simulate external hazards, pilot inactivity or exacerbating inputs can represent inappropriate crew response, spiral dive, stall or varying attitude/velocity can serve as a vehicle upset, and the scenario can be run for takeoff, cruise or approach conditions, i.e., a given scenario can be simulated in many different ways.

The selected combinations consistent with the reality of aviation, a factor that can contribute to flight safety. Furthermore, during the data simulation, several variables and conditions of the aircraft were detected, which allows us to present a part that became the focus of this article is the simulation of tasks with different scenarios. Therefore, it is extremely important that equal conditions are offered to pilots and thus tasks are carried out coherently. The simulations were carried out with one pilot at a time and, aiming to maintain maximum constancy of all external variables, a procedure was established and followed for each pilot. It consisted of pre-flight briefing, task execution and maneuver debriefing and assignment of Cooper-Harper classification.

The initial conditions of each task (i.e., aerodynamic configuration, altitude, airspeed, bank angle, and heading) were established to ensure reproducible of the manoeuvres and equal conditions for every pilot. During the briefing it's necessary to guaranteed a precise definition of the beginning and ending of each task, which was important to clarify the period to which the qualitative assessment referred and for a precise monitoring and recording of the quantitative data.

The characteristics of the research required pilots experienced in engineering flight testsand assessment of management qualities, which, unfortunately, correspond to a very restricted group. With the help of Embraer, seven experienced test pilots were invited to participate in the simulations. In the Table 1 it is possible to see a summary of the pilots' experience and history.

Pilot	Initial Formation	Type Ratings	Flight Hours
1	Civilian	A310; A319; A320; A330; A350 B737-200/-300/-400; F-100; MD-11	19000
2	Military	E120; E145; E170; E390; E530	6000
3	Military	E145; E170; E390; E530	7500
4	Civilian	E145; E170; E390	6000
5	Military	AMX; E110; E120; E145; E170; E190; L500	7500
6	Military	AT-26; C-95; E145; E170; E190; F-5; HS-125; L500	7500
7	Civilian	E50P; E55P; E120; E145; E170; E190; E550	5800
Mean (SD)			8470 (4360)

Table 1 – Pilot Information

2.1 Scenario 1

This scenario, simulates a single engine failure, representing 3.97% of accidents and 3.12% of fatalities. The most critical condition for testing is a total loss of thrust, irrespective of which engine fails, during takeoff when maximum engine power is required. Scenario 1 was straightforwardly reproduced in the simulation campaign by shutting down engine 2, which was selected arbitrarily.

2.2 Scenario 2

This scenario, simulates the failure of a control surface actuator and accounts for 1.59% of the accidents and 1.49% of the fatalities. For simulation purposes, the total loss of any given primary control surface significantly increases the chances of an accident, and does not represent the redundancies of a fourth-generation aircraft well; therefore, a 75reduction was selected for the experiments.

2.3 Scenario 3

This scenario simulates an unresponsive engine combined with inappropriate pilot actions, resulting in aircraft upset conditions such as abnormal angular rates and/or stall. It represents 2.38% of accidents and 1.74% of fatalities. For simulation purposes, the unresponsive engine precursor was modeled by locking engine 2's output power at 50% of the total available thrust and disabling communication between its corresponding throttle lever and FlightGear. Regardless of the throttle lever position, engine 2 consistently provided half of its maximum power. Pilots were not permitted to shut down the engine. The "inappropriate pilot actions" precursor was simulated by applying exaggerated and delayed control inputs, represented a reduction in the sidestick sampling rate from 60 Hz to 3 Hz.

2.4 Scenario 4

This scenario, simulates an icing impairment leading to abnormal attitudes and/or stall. It accounts for 14.29% of accidents and 8.74% of fatalities. The mechanisms of ice accretion are complex and highly specific to each aircraft type, necessitating wind tunnel experiments to develop representative models of such conditions. FlightGear lacks icing simulation options, and there are no studies in the literature on ice accretion for the Boeing 777. However, it is generally known that icing reduces the effectiveness of lifting and control surfaces, as well as engines.

The "icing" aircraft model used in this research may not fully represent the behavior of a Boeing 777-200ER during ice events, but the modifications made were based on the suggestions of Belcastro and the known effects of ice accretion. Inappropriate pilot actions are anticipated during ice events, as pilots are not typically trained for these conditions. This precursor was addressed in the simulations by introducing a delay between sidestick inputs and vehicle response. Specifically, this delay was implemented by reducing the sampling rate of the EB2's sidestick position from 60 Hz to 1 Hz.

Scenario ID	Summary	Precursors
01	Single engine failure	100% thrust reduction in engine 2
02	Loss of control surface effectiveness together with exacerbated crew inputs	Loss of 75% of the elevator effectiveness Exacerbated crew control inputs
03	Unresponsive engine together with exacerbated and delayed crew inputs	Engine 2 unresponsive and locked at 50% of the total available thrust Exacerbated and delayed crew control inputs
04	Icing impairment together with delayed crew inputs	Icing accumulation (sooner stall, less effective control surfaces, thrust asymmetry) Delayed crew control inputs

Table 2 – Scenarios and their Precursors

Four out of the 13 scenarios initially selected met the simulation criteria, and account for 24.61% of the LOC-I accidents and 17.57% of the associated 6087 casualties analysed by Belcastro, C. and Foster, J.,38 which are representative numbers. Towards avoiding nomenclature confusion, the four remaining scenarios were renumbered, as shown in Table 4, along with their short descriptions.

3. Results and Discussion

According to the methodology proposed by Cooper and Harper, the rating must express the evaluation of an aircraft in one given task. So, it is not expected great variations when a given task is proposed to different pilots on a same aircraft.

However, the results showed a big variance, for example in the two diagrams showed in Figure 2, a same task, in the same scenario for a same aircraft results, for different pilots, CHRs from 2 to 9. Demonstrating that the pilot subjectivity is an important issue to be considered, and can take a strong impact on the evaluation, affecting the certification of an aircraft with marginal handling quality.

The experiments were performed in the EB2 research flight simulator at the São Carlos School of Engineering (EESC-USP), in Brazil. The simulator features a 6DOF (Degree of Freedom) hydraulic hexapod motion system (Stewart platform). For safety reasons, saturation thresholds were programmed to limiting the motion of the platform.

3.1 Scenario 1

Task 1

Taking into account that this scenario consists of the failure of an engine, the return was carried out with the engine running and at high speed. Regarding the workload associated with the maneuver, the pilots claimed that the physical workload was “small” and the mental workload varied from “small” to “moderate”. All pilots assigned CHRs in the "normal" controllability category.

Degree of Freedom	Displacement
Longitudinal	-370 mm to 400 mm
Lateral	-400 mm to 400 mm
Vertical	-275 mm to 275 mm
Roll	-15° to 15°
Pitch	-15° to 15°
Yaw	-15° to 15°

Table 3 – EB2 motion limits (relative to the central position of the platform)

Task 2

According to the recorded data, all pilots adopted similar control strategies and two pilots expressed surprise at the increase in power required by the aircraft to maintain its speed during the maneuver.
Four pilots felt that the airspeed performance requirement was "difficult" to meet, and comments suggested that airspeed control was the reason two pilots rated the mental work load level as "moderate". From the perspective of the Cooper-Harper and Quantitative Loss of Control Criteria (QLC) scales, the event was of “normal” controllability for all pilots.

Task 3

In this maneuver, the aircraft presented a high level of energy and a quick turn was made over the running engine. The pilots came to the consensus that controlling the aircraft was “easy.” however, different strategies were identified and this difference may originate from the pilot’s individual interpretations of the task request.
Regarding workload, all pilots understood that both physical and mental participation were “small”, except for ‘Pilot 5’, who classified the latter as “moderate”. The Copper-Harper scale and QLC classified the event as having “normal” controllability for all the pilots.

Figure 2 – Cooper-Harper Rating classified by 6 different pilots.

3.2 Scenario 2

Task 1

The classification of physical and mental workload levels was between “low” and “moderate”. All airmen assigned the CHRs the “normal” controllability category, and the same classification was made by the QLC.

Task 2

The pilots stated that the inclination angle required in the task required inclination inputs to maintain altitude, however, the “slow” longitudinal response, combined with an “abrupt” or even “excessive” lateral response, revealed that the general difficulty of the task it was "reasonable".

The pilots classified physical and mental workload levels between “small” and “moderate”. The event was "normal" for both qualitative and quantitative evaluation methods (’Pilot 2’ and ’Pilot 6’ assigned the lowest CHRs and did not cross any QLC envelope, including applying a different control strategy in pitch when compared to other pilots).

Task 3

The controllability of the event was "normal", both from the perspective of the Cooper-Harper scale and the QLC; however, large discrepancies were observed among the assigned CHRs (from 2 to 7), which was not the case for the number of crossed envelopes (for all pilots no envelope was crossed).

Task 4

Overall, pilots rated the difficulty of this task as “fair” to “high” and ‘Pilot 5’, which applied significantly lower amplitudes. All pilots assigned CHRs in the "normal" controllability category, and most of them crossed a QLC envelope; therefore, the event was “normal” also for the quantitative method.

Task 5

According to the Instrument Landing System approach maneuver, the pilots understood that controlling the pitch and roll attitudes were of equal difficulty ("reasonable" to "high"), despite the problems of capturing the glideslope having been considerably more frequent than localized capture. Regarding controllability categories, both methods classified the maneuver as “normal”.

The only exception was ‘Pilot 5’, he was the only pilot to report “high” physical and mental workloads during the task. Comments revealed that the pilot had control of the aircraft; however, it did not follow the inputs and, consequently, the obligation was to constantly reflect on the aircraft’s response; as a result, the workload exceeded the tolerable limit. According to the comments, ‘Pilot 5’ assessed the condition differently than other aviators due to perceptions of workload rather than the use of a specific control strategy.

3.3 Scenario 3

Task 1

An Instrument Landing System approach was also performed in Scenario 3, and most of pilots reported a "fair" to "hard" control of the pitch attitude. Overall, the controllability of the aircraft was classified as "normal", except for one of the pilots, who understood the workload was beyond tolerable, and Pilot 1, who crossed two QLC envelopes and for CHR in this task the avarange was from 5 to 9.

Task 2

Three pilots classified the workload between “moderate” and “high”, while the other three reported that physical participation was “small”. Differences in perception directly affect the CHRs, since the first and second groups tended to assign classifications in the categories “borderline condition and “normal condition”, respectively.

Task 3

The tasks 2 and 3 were similar, but the aircraft had a lower energy level in the latter. As a result of applying the Cooper-Harper scale, only one of the pilots classified in the “borderline condition” category. Although for another 2 pilots the workload was beyond tolerable, controllability was not threatened and the event was “normal” for the remaining pilots.

Task 4

The constant need to act on the controls during this task was identified, which became essential for the pilots to classify the physical and mental workloads between “intermediate” and “high”. The qualitative method showed

Figure 3 – Cooper-Harper Rating classified by 6 different pilots.

good convergence in classifying controllability as “normal” (only ‘Pilot 1’ assigned CHR in the “borderline condition” category, but it was not possible identify the reasons for this in your comments), while the quantitative method was more divergent (two pilots in each controllability category).

Task 5

The pilots adopted a control strategy based on the use of smaller aircraft control amplitudes compared to Scenario 3 - Task 4, which was a direct consequence of the aircraft’s energy level.
The event was “normal” according to the Copper-Harper scale. However, in general, CHRs were higher in Task 4, as some pilots felt they faced a workload beyond what was tolerable. The comments revealed that such an increase in workload was due to airspeed control, since the higher speed at the beginning of this Task required more frequent actuation by the pilot on engine 1 to meet the non-overspeed performance criterion.

Figure 4 – Cooper-Harper Rating classified by 6 different pilots.

3.4 Scenario 4

Task 1

Os CHR atribuídos ao final da manobra pareciam refletir a opinião dos pilotos sobre um possível PIO; entretanto, com exceção de ‘p05’, o evento foi de controlabilidade “normal” tanto para os métodos de classificação qualitativos quanto para os quantitativos.

Task 2

The CHRs in this Task 1 appeared to be linked to a pilot’s impression of a possible PIO, as the highest ratings were assigned by two pilots. In general, for both qualitative and quantitative classification methods, the event was “normal”; the only exception was a pilot.

Task 3
From the Cooper-Harper scale and QLC perspectives, the event was in the “normal” controllability category for most drivers. there was only an outlier among the pilots, for whom the quantitative method classified the maneuver as a “borderline condition.

Task 4
Despite expected differences between tasks in terms of control performance, pilots reported an increase in mental workload in Task 4 (trending towards “high”, as opposed to “moderate” in Task 3). Under normal circumstances, such information is promptly feedback so that pilots can correct it, however, in this case there is a delay between entries and exits in Scenario 04, a fact that contributed to surprising the pilots. Although both CHR and QLC captured such degradation for some drivers, most of them still classified the event as “normal”.

Task 5
The most pilots agreed that there was a tendency for pitch and for ‘Pilot 5’ it was light, but the others emphasized that it was expressive and the reason for using special control techniques that aimed to reduce the input gain. For this task, one of the pilots demonstrated mentally overwhelmed and simply stopped the maneuver.
This high workload demand was noted by the CHRs as they classified the event as a “borderline condition” for the majority of pilots. From the QLC perspective, however, the maneuver it was "normal".

Task 6
Although pilots worked less on the controls in Task 5, they reported a "high" workload during it, and the reason for this was the "significant" tendency of aircraft deviate to a longitudinal PIO.
Finally, with regard to the event’s controllability rating, two drivers assigned CHRs in the "borderline condition" category, while the other four gave ratings in the "normal" category. About QLC point of view, for three pilots the maneuver was “normal”, and for the another three, the “LOC-I”.

Task 7
Regarding the controllability classification of the event, it was “normal” for three pilots and “LOC-I borderline condition” for one of them according to the qualitative and quantitative criteria. Another pilot also presented a “LOC-I boundary condition” for the qualitative approach, but a “normal” event for the quantitative method. Data is missing for one of the pilots in this maneuver.

Task 8
In Task 7, the aircraft was faster and turn was over engine 2, which was a more natural attitude in the conditions of Scenario 04, and meant that less energy was required to execute the turn. For both the Cooper-Harper scale and the QLC, the event was of “normal” controllability for all the pilots.

Task 9
All pilots reported that this maneuver was “difficult” in all aspects and required “high” mental workload and one of the reasons is that an ILS approach is a naturally demanding task, as the margins are small.
According to all pilots, the motivation for using special control techniques was the need for better control conditions.
Regarding the qualitative classification of controllability of the event, 3 pilots understood that it was a “LOC-I”, since they were unable to even try to effectively interpret the aircraft’s behavior. CHRs ranged from 5 to 10 and there was not enough data for ‘Pilot 4‘.

Task 10
There was a difference in control strategies, which seemed to find a correspondence in the CHRs, since handling qualities were better for pilots who frequently used the throttle levers during the maneuver.

Task 11
In the end, the Cooper-Harper scale dictated that the event was "normal" (the exception was ’Pilot 2’), but the QLC classified it as "normal" for ’Pilot 5’, a "LOC-I borderline condition" for ’Pilot 1’, ’Pilot 3’ and ’Pilot 6’, and a "LOC-I" for ’Pilot 2’ and ’Pilot 4’.

Task 12

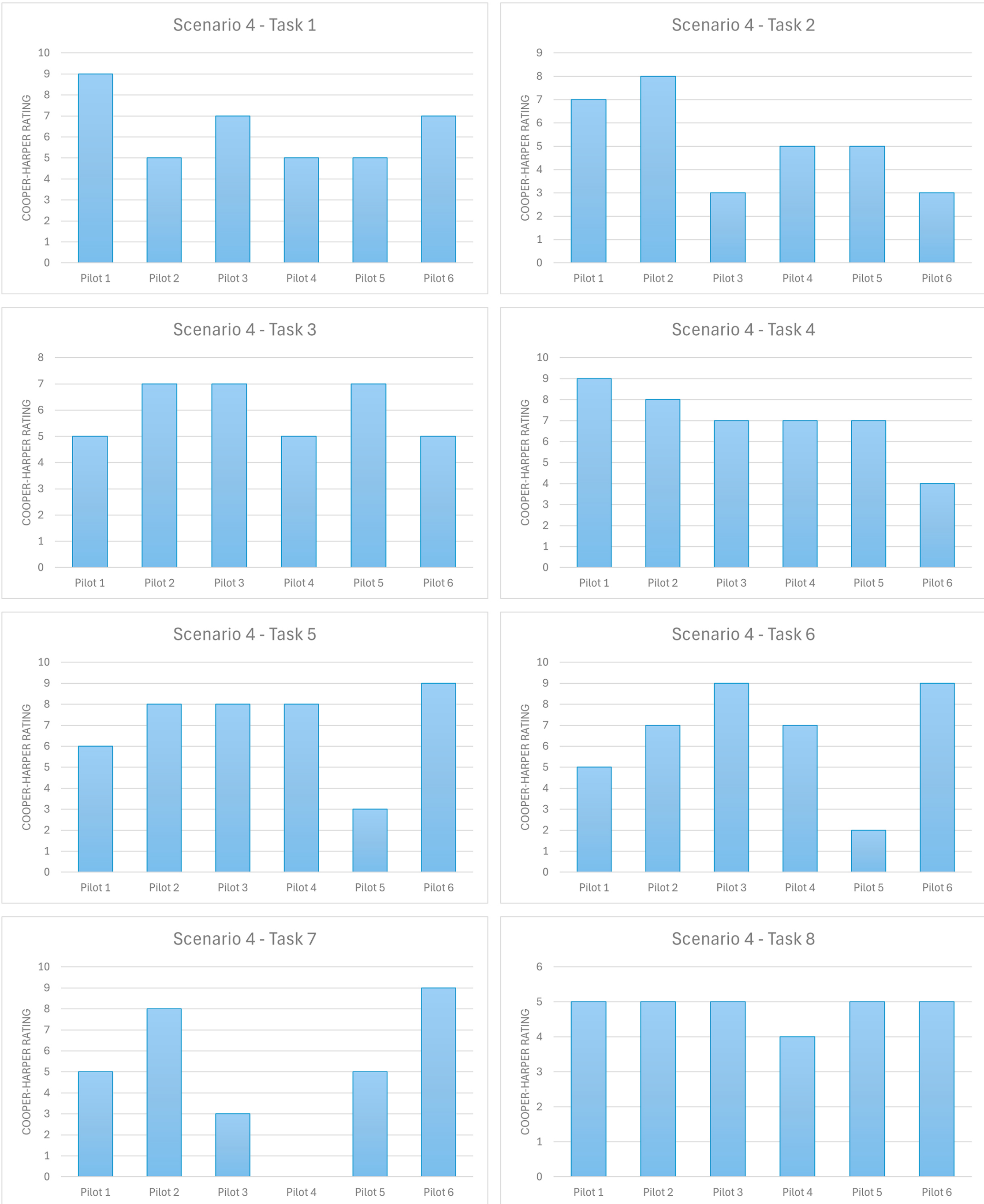

Figure 5 – Cooper-Harper Rating classified by 6 different pilots.

Para a maioria dos pilotos, a carga geral de trabalho da tarefa ficou entre “moderada” e “alta”, e eles atribuíram CHRs dentro da categoria de controlabilidade “normal” (a única exceção foi ‘Pilot 1’, que atribuiu um CHR na categoria “condição limítrofe" categoria).

Task 13
Compared to Task 12, the higher aircraft energy in Task 13, as expected, reduced the average however, the performance of ‘Pilot 1’ and ‘Pilot 3’ was considerably different from the strategy adopted by the other pilots, as evidenced by the significant standard deviations of the mentioned parameters and inevitably increased their workload. However, the CHRs assigned by all other pilots were within the “normal” controllability range, even ‘Pilot 3‘ had the highest CHR.

From all of the aforementioned along this section, we can conclude that pilots were given the opportunity to explore and try different control strategies in the experiments, since, in some occasions, we identified the use of different strategies by groups of pilots (e.g, in Scenario 4 - Task 11).
Furthermore, we also concluded that the simulation scenarios and execution of each task, as expected, actually impacted the pilots’ workload and, consequently, the assigned CHRs and the number of QLC envelopes crossed. However, through qualitative analyses, it was not possible to identify the form and extent of such an impact on controllability classification methods, mainly because there were occasions when pilots used the same control strategy, but the methods classified the event in a different way, different for each of them. It is expected that this is due to the interpretative options of the Cooper-Harper scale, and the fact that maneuvers normally bring the aircraft closer to the edges of some QLC envelopes, that is, quantitative references.

In this context, this paper brings light to an important discussion: is the evaluation using the Handling Qualities Rating Scale a reliable tool for aircraft certification? Is it possible to develop a tool, based on Data Science and Artificial Intelligence, to classify the handling quality of an aircraft?

4. Conclusions

The maneuvering quality and control of an aircraft changes over time, i.e. they can improve/degrade, which leads us to realize that there are nuances to arriving at a result and that this is not something good or bad. While in the original Cooper-Harper scale proposal pilots had ten ratings to express their opinion on the handling qualities of an aircraft, when we reorganized it there were only three categories to represent pilots’ understanding of an aircraft’s controllability. Consequently, the rearrangement of the CHR meant that the nuances of controllability could not be captured, suggesting that controllability is much more of a “discrete” variable than a “continuous” variable. Most of the tasks performed were more demanding in either the longitudinal or lateral axis, however, the CHR assigned at the end of each manoeuvre them assessed the "overall" controllability of the aircraft, it did not represent neither of the axes.

The pilots evaluated their degree of compensation during the handling qualities/controllability assessment process, and such an evaluation depended on several variables, among which the sensation of motion and the feedback provided by the controls. In that regard, the EB2 research simulator has limitations that possibly softened the dynamics characteristics of the aircraft, and consequently misled pilots in their evaluation, culminating in the assignment of CHRs that tended towards the "normal controllability category". Although pilots had not recurrently complained about the simulator during the experiments, it is foreseen that the saturation limits that impeded the platform to move beyond 15 °in pitch and roll, and the absence of control force feedback significantly reduced their sensation of control effort, contributing to the inequality of controllability data distribution. It is still relevant to mention that concerns of such an order are not exclusive to EB2 since they also apply to flight simulators used for real pilot training, inasmuch as the aviation authorities push the industry to improve the emulation of motion in simulators and make them more realistic.

According to the presented graphs, I highlight two examples of significant discrepancies among the pilots concerning the qualitative method of the Cooper Harper Rating. This clearly indicates the need to reassess the highly subjective nature of this method, which compromises the accuracy of the obtained data.

Figure 6 – Cooper-Harper Rating classified by 6 different pilots.

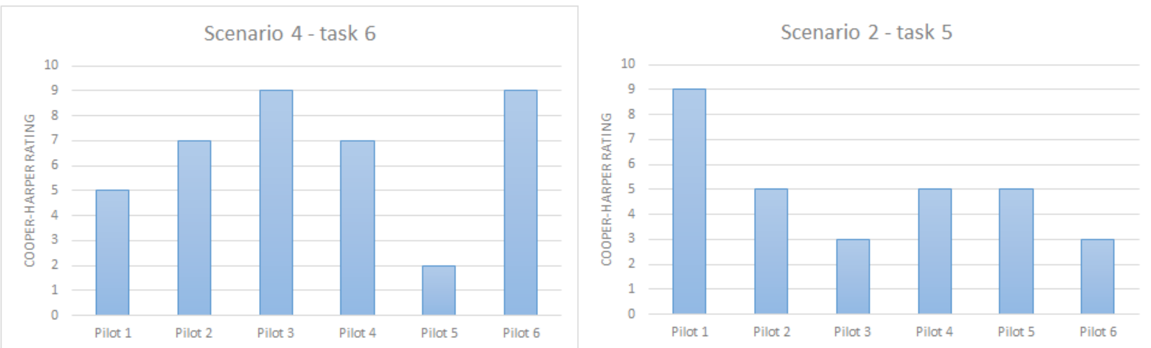

Figure 7 – Cooper-Harper Rating classified by 6 different pilots for two of the tasks performed.

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] Cooper G, Harper Jr R *The use of pilot rating in the evaluation of aircraft handling qualities*. Washington DC: National Aeronautics and Space Administration, 1969.

[2] David G. Mitchell *Fifty Years of the Cooper-Harper Scale*, AIAA Scitech 2019 Forum

[3] Milward, Thomas Oliver and Bromfield, Mike and Horri, Nadjim and Ali, Rashid and Scott, Steve *Multi-point Angle of Attack Sensing for Avoidance of Loss of Control in Flight: AIAA SCITECH 2019*, 2018

[4] Belcastro C. Validation of safety-critical systems for aircraft loss-of-control prevention and recovery. *AIAA Guidance, Navigation, and Control Conference*, ugust. American Institute of Aeronautics and Astronautics, 2012.