Comparing Active and Passive Just Noticeable Difference Thresholds for Stall Abruptness in Symmetric Stall

Bootsma, S.; de Visser, C.C.; Pool, D.M.

DOI
10.2514/6.2025-0975

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the AIAA SCITECH 2025 Forum

AIAA SciTech Forum
6-10 January 2025, Orlando, FL
AIAA SCITECH 2025 Forum

10.2514/6.2025-0975

Comparing Active and Passive Just Noticeable Difference Thresholds for Stall Abruptness in Symmetric Stall

Sybren Bootsma*, Coen C. de Visser† and Daan M. Pool‡
Delft University of Technology, Delft, Zuid-Holland, The Netherlands

Aerodynamic stall has been a critical factor in recent aircraft crashes, leading to revised regulations for simulator-based stall prevention and recovery training. However, the updated regulations still lack an objectively defined level of accuracy for simulators' stall models that ensures effective pilot training. To help determine this required accuracy, this paper investigates how the Just Noticeable Difference (JND) thresholds for deviations in a stall model's 'stall abruptness' parameter translate from a passive observer setting (typical JND experiments) to an active flying scenario (realistic training task). An experiment was performed in the SIMONA Research Simulator with 16 active pilots, whose sensitivity to stall abruptness variations was measured in both these scenarios. In the passive scenario, pilots' JND thresholds were measured using a staircase procedure in a symmetric stall maneuver flown by a stall autopilot. In the active scenario, the method of constant stimuli was used to determine the JND thresholds when the pilots themselves actively flew the same stall maneuver. The average JND thresholds for the passive and active scenarios were estimated by fitting a psychometric curve to the combined responses of all participants. Overall, the passive JND thresholds for the stall abruptness parameter, with an average Weber fraction of 0.11 ± 0.094**, were lower than those measured in an earlier experiment** (0.16 ± 0.14)**, indicating a higher sensitivity. Furthermore, the psychometric curve of the active experiment was found to lie entirely to the right of the passive psychometric function: the active JND threshold was found to be five times higher than the passive JND threshold. Overall, this indicates a decreased sensitivity to changes in stall abruptness – and hence a reduced demand on its modeling accuracy – when pilots are flying a stall themselves.**

Nomenclature

Abbreviations

2AFC	two-alternative forced choice
2D1U	two-down, 1-up
CDF	cumulative distribution function
FSTD	flight simulation training device
JND	just noticeable difference
RMS	root mean square
PEST	parameter estimation by sequential testing

Roman Symbols

a_1	stall abruptness
a_1^+	upper threshold for a_1
c	chord
C_D	drag coefficient
C_L	lift coefficient
C_l	roll moment coefficient
C_m	pitch moment coefficient
C_n	yaw moment coefficient
C_T	thrust coefficient
C_Y	side force coefficient
$\dot{h}_e$	Vertical speed
$K_{\{q,x,z\}}$	pitch, heave, and surge gain
P	gain
P	probability of correctness at stimuli level
p	statistical significance
p	roll rate
q	pitch rate
r	yaw rate
u, w, z	velocity vectors
V	velocity
X	flow separation point
X_0	steady flow separation point

*M.Sc. student (graduated), Control and Simulation section, Faculty of Aerospace Engineering, P.O. Box 5058, 2600GB Delft, The Netherlands; sybrenbootsma@hotmail.com. Student Member AIAA.

†Associate Professor, Control and Simulation section, Faculty of Aerospace Engineering, P.O. Box 5058, 2600GB Delft, The Netherlands; c.c.devisser@tudelft.nl. Member AIAA.

‡Assistant Professor, Control and Simulation section, Faculty of Aerospace Engineering, P.O. Box 5058, 2600GB Delft, The Netherlands; d.m.pool@tudelft.nl. Associate Fellow AIAA.

Greek Symbols

α	angle of attack
α^*	angle of attack for which $X_0 = 0.5$
β	angle of side slip
δ_a	aileron deflection
δ_e	elevator deflection
δ_r	rudder deflection
$\zeta_{\{q,x,z\}}$	pitch, heave, and surge damping coefficient
θ	pitch angle
μ	mean
σ	variance
τ_1	stall time delay constant
τ_2	stall hysteresis time constant
φ	stimuli level
$\omega_{b_{\{q,x,z\}}}$	pitch, heave, and surge break frequency
$\omega_{n_{\{q,x,z\}}}$	pitch, heave, and surge natural frequency

I. Introduction

Loss of control in-flight (LOC-I) is currently the primary cause of fatal accidents in commercial aviation [1]. These accidents often result from pilots failing to prevent, or recover from, an upset. One of these upsets is the stall, a situation where the critical angle of attack is exceeded, which leads to a sudden loss of lift [2]. Aerodynamic stall has been a primary factor in several recent aircraft crashes [3–5]. Since then, the International Civil Aviation Organization has updated the regulations regarding upset recovery training, which have come into effect in 2019 [2, 6, 7]. Numerous aspects of the upset prevention and recovery training were updated, including the required fidelity of the stall model used in flight simulation training devices (FSTDs). An FSTD is currently certified for stalls by a subject matter expert who evaluates if the used model matches reality well enough [8] or, as mentioned in EASA's regulations [9]: "*for each upset scenario, the recovery manoeuvre can be performed such that the FSTD does not exceed the FSTD training envelope, or when the envelope is exceeded, that the FSTD is within the realms of confidence in the simulation accuracy*".

Previous efforts have validated that stall models match reality sufficiently through pilot-in-the-loop simulations. For example, Schroeder et al. [10] evaluated four different stall models in a Level D B737-800 simulator: an old model from before the 2019 regulations update, an updated model by Boeing for the 2019 regulations, and two models based on scaled wind tunnels tests, computational aerodynamics, and expert opinions. These four stall models were tested by several pilots experienced in flying stalls on the B737. When asked if the stall models could be used for training, the participants "*somewhat agreed*" [10] . Furthermore, they found no significant difference in recovery performance between the models. Finally, they concluded that effective stall models could be developed based on wind tunnel data, computational aerodynamics, and expert pilot opinions.

Other work by Grant et al. [11] used three different models for their comparison. They used a nominal model and compared it with an extreme and a mild model, both of which were derived from the nominal model. The extreme and mild models were defined by changing the aerodynamic parameters. Furthermore, Grant et al. [11] increased the roll-off and stall buffet for the extreme model, and decreased their intensities for the mild model. They found no significant difference in recovery performance of participants between the different models, consistent with [10]. Moreover, their participants indicated that each of the models developed were acceptable for stall training.

A first explicit and quantitative analysis of pilots' sensitivity was performed by Cunningham et al. [12], who evaluated the sensitivity to changes in several characteristics of their stall model. These factors included, among others, stall asymmetries, control effectiveness, and dynamic stability. They asked participants to rate the significance of the changes with respect to these factors on a scale of 0-9, where 0 meant no differences and 9 represented large differences. Their participants indicated that the stick pusher and stall asymmetries had the most significant changes in the simulation, while the dynamic stability showed the least.

A similar effort was performed by Smets et al.[13] and Imbrechts et al. [14], who measured the Just Noticeable Difference (JND) thresholds for key parameters of their stall model and stall buffet model, respectively. This was done through simulator experiments using a staircase procedure where the baseline model was compared to a model that contained a deliberately adjusted parameter value. Overall, their results indicate that pilots may not reliably notice stall model parameter offsets up to 10-30% [15]. However, in these experiments the participants were not in control of the approach-to-stall and stall recovery: a specifically designed stall autopilot [13] flew the maneuver. Hence, participants could use their full attention for detecting differences between the baseline and adjusted models.

This leaves the question what happens to pilots' JND thresholds when they fly the aircraft themselves. For this, other research at TU Delft [16, 17] provides insights. Hosman and Van der Vaart [16] investigated how passive vestibular motion perception thresholds change when additional mental workload is added. They found an increase in threshold of 26% to 266%, depending on the added mental workload. Valente Pais et al. [17] looked into the absolute threshold in pitch motion during an active control task and a passive observing task and found that the absolute threshold was 60%

higher in the active control experiment. Both works indicate that a higher threshold is found when additional mental workload or control inputs are required.

Building upon these findings, this research investigates how pilots' JND thresholds differ between a passive observing task as considered in [13, 14] and an active control scenario, for the stall abruptness parameter of a Kirchhoff stall model [18]. This question is answered by performing a pilot-in-the-loop experiment, where the upper JND threshold for both a passive autopilot-controlled scenario and an active pilot-in-command scenario are directly compared. These scenarios will be referred to as 'passive' and 'active' in the remainder of this paper, respectively. Both scenarios were tested in the SIMONA Research Simulator at TU Delft. In the experiment, 16 pilots experienced a symmetric stall maneuver at around 18,000 feet. A Parameter Estimation by Sequential Testing (PEST) staircase procedure [19] was used for the passive scenario, while the method of constant stimuli [20] was used for the active pilot-in-command scenario. For both scenarios, participants compared the baseline stall model to a model with an increased abruptness in the flow separation and were asked to indicate which of the two showed a more abrupt response. Based on the answers given, the JND threshold and psychometric curve were determined for both experiments, which gives direct quantitative insight into changes in pilots' sensitivity between a passive observing task and an active control setting.

This paper is structured as follows. First, Section II provides additional information on the aerodynamic stall model used in this research, as well as the changes made to the model to allow active pilot control. Section III provides the hypotheses for this research and describes the experiment set up used to determine the active and passive thresholds for changes in stall abruptness parameter of the Kirchhoff stall model, followed by the results of the experiment in Section IV. Finally, the results are discussed in Section V, followed by the main conclusions in Section VI.

II. Aerodynamic Stall Model

A. Stall Model based on Kirchhoff's Theory of Flow Separation

The aerodynamic stall model used in this experiment has been developed by earlier research at TU Delft [18, 21] and is based on Kirchhoff's Theory of Flow Separation, as first described by Fischenberg [22]. This model is centered around the calculation of a flow separation point state variable, X. The flow separation point is defined to be 1 when the flow is fully attached to the wing and 0 when the flow is fully separated. X is dependent on many factors, such as the hysteresis factor, which captures the circulation and boundary layer effects through Kirchhoff's theory. By combining this with the Wagner or Theodorsen function to capture the unsteady aerodynamics, Fischenberg [22] made the flow separation point time-dependent. Finally, X can be estimated from the steady flow separation point, when combined with the previous aspects. This ultimately results in the nonlinear ordinary differential equation that can be used to dynamically model the flow separation, as can be seen in Equation (1). For more information regarding the model, see Van Ingen et al. [18] and Fischenberg [22].

$$\tau_1 \frac{dX}{dt} + X = \frac{1}{2}\{1 - \tanh(a_1(\alpha - \tau_2\dot{\alpha} - \alpha^*))\} \tag{1}$$

In Equation (1), the different parameters all capture different characteristics of the stall. a_1 $[-]$ is indicative of the stall abruptness, where a higher value means a more abrupt drop in lift and a more sudden and dramatic flow separation. This is the parameter that will be varied to measure the threshold for changes in stall abruptness in this paper. Furthermore, α^* $[rad]$ is the angle of attack for which the flow separation point $X = 0.5$. τ_1 $[s]$ and τ_2 $[s]$ are both time constants that capture the transient and hysteresis effects of the stall, respectively [13, 21–23]. The flow separation point X is used in the longitudinal aircraft model of the Cessna Citation II, the research aircraft of the Faculty of Aerospace Engineering at TU Delft*. This model is obtained from Van Ingen et al. [18], with the longitudinal model as given by Eq. (2) to (4), with the exception of the C_{m_q} term indicated in red which will be discussed in Section II.B. The coefficients are given in Table 1. The lateral model used in this research is given by Van Ingen et al. [18] as well. Furthermore, the adaptive, X-dependent stall buffet model as described by Van Horssen et al. [21] is used.

$$C_L = C_{L_0} + C_{L_\alpha}\left(\frac{1+\sqrt{X}}{2}\right)^2 \alpha + C_{L_{\alpha^2}}(\alpha - 6)_+^2 \tag{2}$$

$$C_D = C_{D_0} + C_{D_\alpha}\alpha + C_{D_{\delta_e}}\delta_e + C_{D_X}(1 - X) + C_{D_{C_T}}C_T \tag{3}$$

*https://cs.lr.tudelft.nl/citation/

$$C_m = C_{m_0} + C_{m_\alpha}\alpha + C_{m_q}\frac{qc}{V} + C_{m_{X\delta_e}}\max(\frac{1}{2}, X)\delta_e + C_{m_{C_T}}C_T \tag{4}$$

Table 1 Baseline stall model parameters from [18].

Name	Value	Unit	Name	Value	Unit	Name	Value	Unit
a_1	27.6711	[-]	C_{D_X}	0.0732	[-]	C_{l_r}	0.1412	[-]
α^*	0.2084	[rad]	$C_{D_{C_T}}$	0.3788	[-]	$C_{l_{\delta_a}}$	-0.0853	[-]
τ_1	0.2547	[s]	C_{Y_0}	0.0032	[-]	C_{m_0}	0.0183	[-]
τ_2	0.0176	[s]	C_{Y_β}	-0.5222	[-]	C_{m_α}	-0.5683	[-]
C_{L_0}	0.1758	[-]	C_{Y_p}	-0.5000	[-]	$C_{m_{X\delta_e}}$	-1.0230	[-]
C_{L_α}	4.6605	[-]	C_{Y_r}	0.8971	[-]	$C_{m_{C_T}}$	0.1443	[-]
$C_{L_{\alpha^2}}$	10.7753	[-]	$C_{Y_{\delta_a}}$	-0.2932	[-]	C_{n_0}	0.0013	[-]
C_{D_0}	0.0046	[-]	C_{l_0}	-0.0017	[-]	C_{n_β}	0.0804	[-]
C_{D_α}	0.2372	[-]	C_{l_β}	-0.0454	[-]	C_{n_r}	-0.0496	[-]
$C_{D_{\delta_e}}$	-0.1857	[-]	C_{l_p}	-0.1340	[-]	$C_{n_{\delta_r}}$	0.0492	[-]

B. Additional Pitch Damping

During the initial testing phase of the experiment, it was found that, when manually flying the stall model, the pitch rate q was not properly damped, despite the Cessna Citation II having a strongly damped short-period [24]. Consequently, the eigenvalues of the $q - \delta_e$ model in the stall regime were determined by reducing the states and finding the eigenvalues in MATLAB. From this, a positive pair of complex eigenvalues was found, with a value of $2.42 \cdot 10^{-3} \pm 0.192i$. The Cessna Citation II should have a damped short-period with eigenvalues of $-3.9161 \cdot 10^{-2} \pm 3.7971 \cdot 10^{-2}i$ in nominal flight conditions [24].

Consequently, a pitch damping coefficient C_{m_q} was added to the pitch moment equation, since the model by Van Ingen et al. [18] did not contain this term, resulting in Equation (4) with the additional C_{m_q} shown in red. Initially, a C_{m_q} value was estimated based on the work by Van den Hoek et al. [25], who identified a nominal flight envelope model. Their work allowed to interpolate a C_{m_q} value based on altitude and Mach number, resulting in $C_{m_q} = -8.1826$. This updated model was verified through pilot-in-the-loop evaluations with an experienced Cessna Citation II test pilot. These experiments verified that the damping that was included was still not sufficient, see Figure 1. Although the addition of C_{m_q} damped the oscillatory pitch rate, the damping was not as strong as the behavior found in the Cessna Citation II, according to the test pilot. Consequently, the value of C_{m_q} was further increased based on offline simulations, which can be seen in Figure 1. Here, the control inputs of the pilot-in-the-loop test (with $C_{m_q} = -8.1826$) were used to analyze the effects of an increase in C_{m_q}. Figure 1 shows that $C_{m_q} = -22$ sufficiently damps the found behavior. The updated model was again verified with pilot-in-the-loop evaluations, which confirmed that the updated model represented the pitch behavior of the Cessna Citation II in the approach to stall and post-stall flight.

The short-period eigenvalues of the updated model in the stall regime were found to be $-7.19 \cdot 10^{-3} \pm 0.142i$. Despite the fact that these values are half an order of magnitude away from the values given by Mulder et al. [24] in nominal flight conditions, the confirmation through the pilot-in-the-loop simulations is deemed acceptable for this research. Further integration of the C_{m_q} within the stall model is outside the scope of this paper and left as a crucial recommendation for future research.

As a final verification step, a comparison of a passive (autopilot-controlled) stall maneuver between the updated model and the model used by Smets et al. [13] was made. This was done to ensure that the differences between the models did not result in different threshold values. The results for the comparison between the pitch angle for the two models can be seen in Figure 2. Here, it can be seen that the difference in pitch between the baseline and the JND threshold found by Smets et al. [13] is similar to the difference between the baseline and the JND threshold setting for a_1 with the updated model. Therefore, it is assumed that the thresholds found in this research can be compared to the passive threshold found by Smets et al. [13].

C. Stall autopilot sensitivity analysis

For the active experiment, pilots own variations in how the stall maneuver is flown will affect the reliability with which changes in the model's stall abruptness can be detected. To be able to compare the JND thresholds for both the

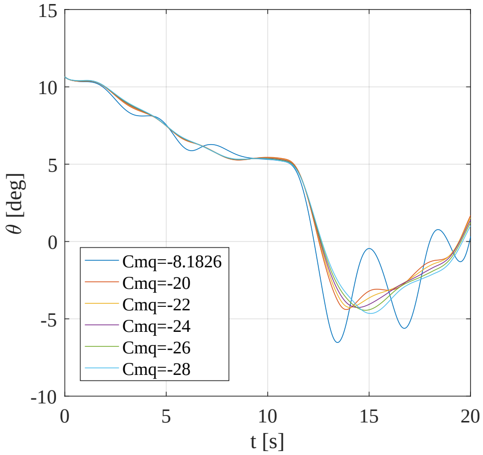


Fig. 1 Pitch angle data from the pilot-in-the-loop experiment in SIMONA with different C_{m_q} values using the control inputs from the experiment.

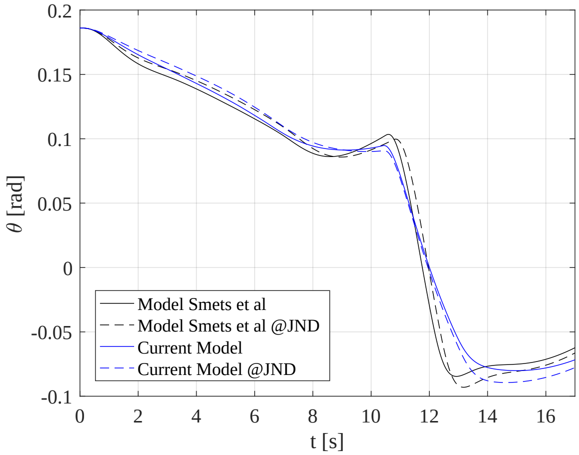


Fig. 2 Comparison of the simulated pitch angle response with updated C_{m_q} and the match to the model output of [13].

active and passive scenarios, both scenarios must be flown as similarly as possible. On the other hand, letting participants focus too much on following an exact flight path can lead to distractions from the actual experiment. Consequently, a balance needed to be found between flying both scenarios similarly versus giving participants the freedom to fly as they are trained.

To determine which part of the maneuver needed to be restricted and could be left up to the participants, an offline sensitivity analysis was performed. The stall autopilot decision tree as used by Smets et al. [13], see Figure 3 and the source paper for more details, was taken as a reference and a sensitivity analysis was performed on the key parameters of the autopilot. This mimicked the differences expected in the control behavior of real participants. The key parameters of the stall autopilot that were varied are:

- The threshold $\alpha_{\text{threshold}}$ at which the recovery procedure is initiated: baseline $\alpha_{\text{threshold}} = 16.04°$
- The reference angle θ_{ref} during the stall recovery phase: baseline $\theta_{\text{ref}} = -0.5°$
- The controller gain P_θ of the reference angle θ_{ref}: baseline $P_\theta = 0.4\ [-]$
- The threshold on $\dot{h_e}$ when full thrust is applied: baseline $\dot{h_e} = -18\ [m/s]$
- The threshold on V_{tas} when the reference angle switches back to 10° to go back to the original flight path: baseline $V_{tas} = 86\ [m/s]$

The threshold for starting the recovery procedure $\alpha_{\text{threshold}}$ was used as a starting point for the variation. The starting angle of attack was varied initially, to see what a valid range for the sensitivity could be. Values higher than 18.5° were deemed unrealistic, since there is already full flow separation at 16°, and hence this would result in an unrealistic recovery scenario. Furthermore, a $\alpha_{\text{threshold}}$ smaller than 13.5 degrees would not lead to a fully developed stall. Consequently, the variations for the sensitivity analysis were set to vary the parameters with ±15%, resulting in a range of 13.6° to 18.4° for $\alpha_{\text{threshold}}$. As a result, the other parameters were also varied with ±15% for the analysis to allow for a fair comparison between the parameters. The variations in important aircraft states such as pitch, pitch rate, velocity, flow separation point, elevator input, and angle of attack were analyzed. These states were compared to the variations in the states that were found with the upper and lower a_1 thresholds of Smets et al. [13].

It was found that the only parameter that would lead to noticeable variations in the states, is the threshold for starting the recovery procedure, $\alpha_{\text{threshold}}$ [°], which can be seen in Figure 4. This figure shows that the differences in velocity, pitch angle, and elevator input all exceed the upper and lower a_1 thresholds found by Smets et al. [13] and are therefore likely noticeable for participants. The variations of the other parameters resulted in differences in the states that fell within the upper and lower a_1 threshold results. Hence, it is concluded that allowing participants to fly the stall maneuver themselves while restricting the stall entry conditions results in the optimal balance.

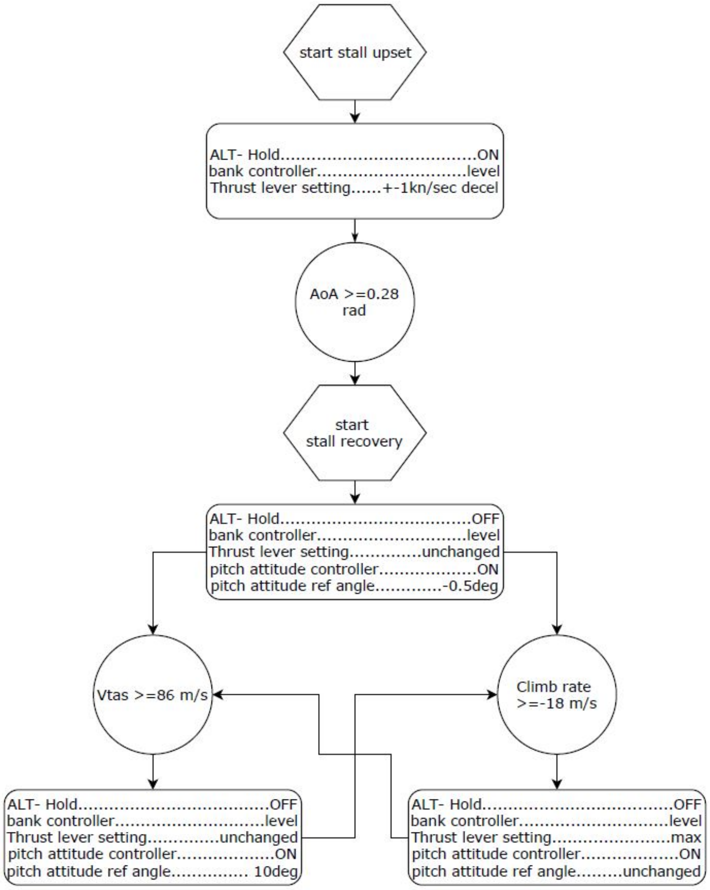

start stall upset
ALT- Hold.....................................ON
bank controller.............................level
Thrust lever setting......+-1kn/sec decel
AoA >=0.28 rad
start stall recovery
ALT- Hold.....................................OFF
bank controller.............................level
Thrust lever setting.............unchanged
pitch attitude controller...................ON
pitch attitude ref angle.............-0.5deg
Vtas >=86 m/s
Climb rate >=-18 m/s
ALT- Hold.....................................OFF
bank controller.............................level
Thrust lever setting.............unchanged
pitch attitude controller...................ON
pitch attitude ref angle............... 10deg
ALT- Hold.....................................OFF
bank controller.............................level
Thrust lever setting.......................max
pitch attitude controller...................ON
pitch attitude ref angle.........unchanged

Fig. 3 Stall autopilot implementation designed by Smets et al. [13].

III. Method

A. Hypotheses

Currently, it is unknown how the JND thresholds for the parameters of the Kirchhoff stall model translate from a passive observer task to an active pilot-in-command task. Based on previous research [13, 15–17] the following two hypotheses were formulated for the experiment:

H1: The active flying JND thresholds for the stall abruptness parameter a_1 will be higher than the corresponding passive JND thresholds. In previous research [16, 17], a comparison between active and passive thresholds is made, albeit that these researches investigated an absolute motion perception threshold rather than a difference threshold. These previous works found that thresholds in an active task were higher than in a passive observer task. Hence, it is expected that this will also be true for stall abruptness JND thresholds.

H2: The upper passive JND threshold for a_1 will be equivalent to the threshold value reported by Smets et al. [13], i.e., $a_1 = 29.72$ or $a_1^+ = 0.074$ expressed as a Weber fraction. Overall, the passive scenario results are expected to be comparable to the previous results of [13], because it is essentially a direct replication of that previous experiment. However, because the current experiment uses a different staircase procedure, which is explained in Section III.B.2, the 70.7% threshold will be obtained instead of the 50% threshold of Smets et al.. Hence, the JND threshold values that are directly derived from participants' individual staircase data cannot be directly compared. However, the average JND thresholds estimated from the psychometric function CDFs fitted to the results of all participants in both experiments are expected to show fully equivalent values.

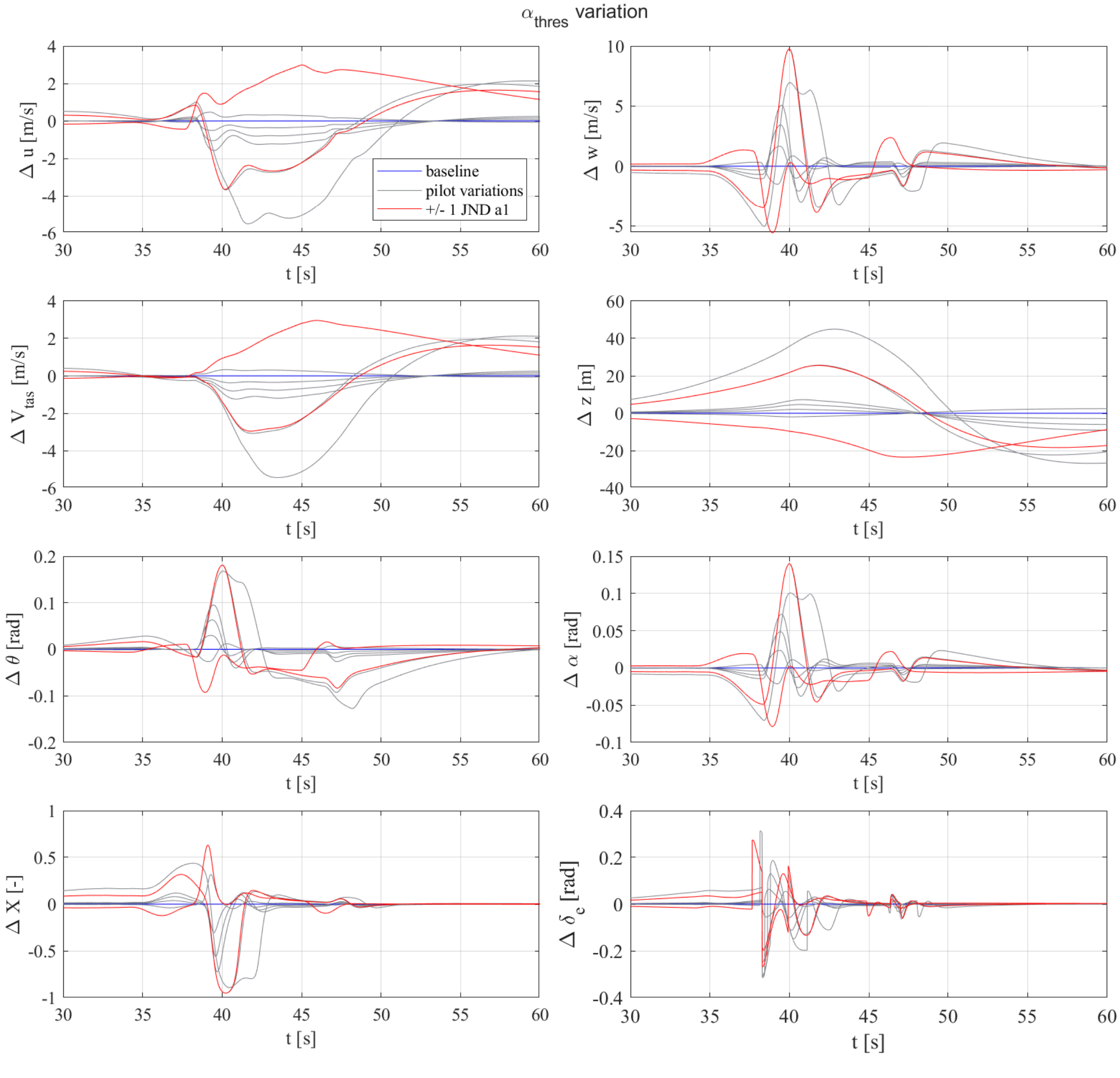


Fig. 4 Difference for relevant states between baseline model and variations in $\alpha_{\text{threshold}}$, as well as the a_1 upper and lower threshold as found by Smets et al [13].

B. Experiment Design and Procedures

The experiment was designed as a within-subjects experiment, where all participants performed both the active (see Section III.B.1) and passive (see Section III.B.2) parts of the experiment. All odd-numbered participants performed the passive experiment first, whereas all even-numbered participants performed the active experiment before the passive scenario. As a result, any potential learning or other order effects were counter-balanced between the passive and active scenarios. Both parts of the experiment were completed in a single session. The session started with a briefing, during which participants signed an informed consent form, filled in the demographic questionnaire, and had the opportunity to ask questions regarding the briefing and safety instructions for the SIMONA simulator. During all measurements, the pilot was the only occupant of the simulator. The true purpose of the experiment was disclosed to the participants after the completion of the full experiment, to prevent any biases from the participants confounding the measured JNDs.

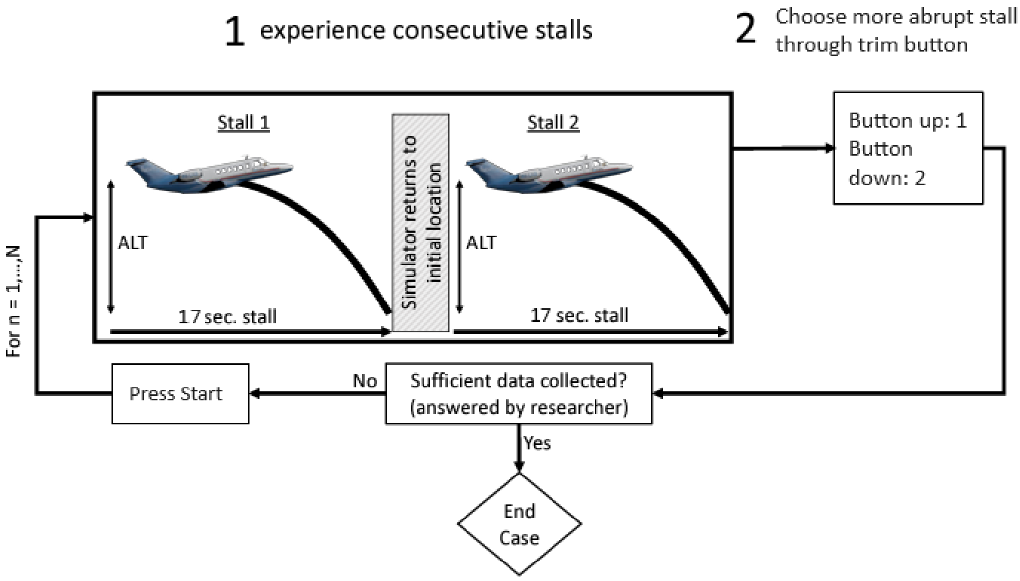

1 experience consecutive stalls
2 Choose more abrupt stall through trim button
Stall 1
Stall 2
ALT
ALT
17 sec. stall
17 sec. stall
Simulator returns to initial location
Button up: 1 Button down: 2
For n = 1,...,N
Press Start
No
Sufficient data collected? (answered by researcher)
Yes
End Case

Fig. 5 Procedures of both the active and passive experiment, adapted from [13].

The active and passive experiment consisted of several consecutive pairwise evaluations, as shown in Figure 5. Participants always experienced two consecutive simulated stalls, with one of them using the baseline value for a_1 and the other a higher adjusted a_1. Participants were instructed to identify the stall with a more abrupt flow separation. This means that the task is a two alternative forced choice (2AFC) task, which has an expectancy of a correct response of 50% by default; i.e., the psychometric curve ranges from 50% to 100%. Therefore, the 50% threshold as found by Smets et al. [13] and the 75% threshold measured in this experiment represent the same (halfway) point on the psychometric curve. A 2AFC procedure is statistically more reliably than the methodology used by [13, 14] – who asked their pilots to indicate if they noticed a difference between the pair of compared stalls, i.e., a yes/no question – as the latter is prone to biases [26]. Participants were told that they would be subjected to the pairwise comparison of two stall maneuvers as shown in Figure 5 an unknown amount of times N, until sufficient data was collected.

The procedures for the active and passive experiment are described in Section III.B.1 and III.B.2, respectively. After the first experiment part was completed, there was a break of approximately 15 minutes before continuing with the second part of the experiment. After the second part, there was a debriefing session during which participants could comment on their experience. They were asked to indicate (if possible) on which cues they had based their answers, see our Discussion section.

Throughout the experiment, participants used the trim switches on the yoke, see Figure 6, to indicate which stall had more abrupt flow separation. Furthermore, participants advanced to the next comparison at their own convenience, by pushing the autopilot disconnect button. Neither of these buttons were used for their original purpose throughout the experiment, meaning that the pilots could not trim the aircraft. The use of these buttons was explained during the briefing. By using these buttons, the staircase was fully automated and participants could determine their own pace in the experiment.

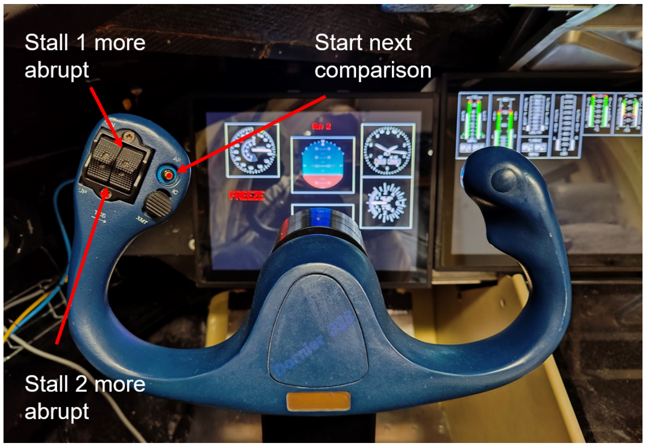


Fig. 6 The buttons used to provide pilot input during the experiment.

Before starting data collection, a training phase was implemented during which participants could familiarize themselves with the scenario. For both the active and passive scenarios, the training consisted only of comparisons between the baseline and the most extreme a_1 setting considered in the experiment. During the active experiment's training phase, extra attention was given to flying the aircraft and ensuring that the participants achieved a consistent flying performance before starting the experiment. Participants were allowed to train until they and the experimenter were confident that they could detect the differences between the baseline and a more abrupt stall.

1. Active Experiment

The active experiment used the method of constant stimuli to assess how the upper a_1 threshold as found by Smets et al. [13] translates to an active flying scenario. As this is currently unknown, the method of constant stimuli provides a first estimate such that in future research, a staircase procedure may be used to efficiently obtain a threshold for each participant. The method of constant stimuli is a rudimentary, non-adaptive method to determine a psychometric function [20]. The method uses 6-9 different, predetermined, levels of the stimulus that are presented several times to participants. Then, by assuming a probability distribution such as the Gaussian distribution in Equation (5), the psychometric function can be created based on the cumulative distribution function (CDF). The corresponding threshold can be found through this psychometric function, by fitting the CDF through the percentage of correct responses of the different stimuli levels. The corresponding JND threshold can be found by, for instance, setting $P(\varphi) = 0.75$ to get the 75% threshold [13].

$$P(\varphi) = 0.5 + 0.5\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\varphi} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}dx \tag{5}$$

In order to assess how the upper threshold found in [13] – i.e., $a_1 = 31.96$ – translates to an active scenario, the tested stimulus levels chosen for the experiment were directly based on this threshold, see Table 2. Three a_1 values above and three values below the JND threshold of [13] are selected as the test conditions, together with the JND threshold setting and the baseline value. Each of the eight a_1 conditions is presented four times to participants, resulting in a total of 32 comparisons for the active experiment.

To balance the experiment conditions, a Latin square is defined with all the tested a_1 settings, see Table 3. The Latin square's rows are labeled A-H. The numbers in the top row represent the testing order, i.e., Latin square row A starts with a baseline comparison, followed by a comparison with +0.25 JND, etc. Because the active experiment tests each condition four times, four of the rows A-H were used per participant. Therefore, another Latin square is generated to balance the eight different test orders A-H evenly across all participants. This Latin square is split between columns 4 and 5, to generate the test sequence for participants 1-8 and 9-16, see Table 4.

During the active experiment, participants had to follow a flight director that would guide them into the stall (see Figure 7a), as was done in the experiment by Cunningham et al. [12]. This was done as the simulations analysis in Section II.C indicated that differences in the approach to stall would quickly lead to differences between the test runs that would be outside of the a_1 threshold reported by Smets et al. [13]. When participants reached full flow separation

Table 2 Active experiment conditions with different a_1 settings.

Condition	$\mathbf{a_1}$ Setting	$\mathbf{a_1}$ Value
1	Baseline	27.6711
2	+ 0.25 JND	28.7427
3	+ 0.5 JND	29.8143
4	+ 0.75 JND	30.8858
5	+ 1.0 JND	31.9574
6	+ 1.5 JND	34.1005
7	+ 2.0 JND	36.2437
8	+ 2.5 JND	38.3869

Table 3 Active experiment balanced condition order Latin Square.

Comparison → **Latin square ↓**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	Baseline	+0.25 JND	+0.5 JND	+1.5 JND	+0.75 JND	+1 JND	+2.5 JND	+2 JND
B	+0.25 JND	+1.5 JND	Baseline	+1 JND	+0.5 JND	+2 JND	+0.75 JND	+2.5 JND
C	+1.5 JND	+1 JND	+0.25 JND	+2 JND	Baseline	+2.5 JND	+0.5 JND	+0.75 JND
D	+1 JND	+2 JND	+1.5 JND	+2.5 JND	+0.25 JND	+0.75 JND	Baseline	+0.5 JND
E	+2 JND	+2.5 JND	+1 JND	+0.75 JND	+1.5 JND	+0.5 JND	+0.25 JND	Baseline
F	+2.5 JND	+0.75 JND	+2 JND	+0.5 JND	+1 JND	Baseline	+1.5 JND	+0.25 JND
G	+0.75 JND	+0.5 JND	+2.5 JND	Baseline	+2 JND	+0.25 JND	+1 JND	+1.5 JND
H	+0.5 JND	Baseline	+0.75 JND	+0.25 JND	+2.5 JND	+1.5 JND	+2 JND	+1 JND

at an angle of attack of $\alpha = 16°$, the flight director disappeared and the "RECOVER" message as shown in Figure 7b would appear. This is the same point as when the stall autopilot would initiate the recovery.

The flight director was programmed to start at 10.65° pitch up and slowly move towards 5.925° pitch up, which matches to the trajectory flown by the stall autopilot from[13]. This pitch up path with decreasing pitch angle ensures that the aircraft enters the stall with a 1 kts/s deceleration. Despite the decreasing pitch angle during this phase, the aircraft's vertical speed is increasing, resulting in an increasing angle of attack and therefore flow separation. Consequently, the participant will reach an angle of attack of $\alpha = 16°$ within 10.5 seconds, which leads to a pitch drop of 0.45 $°/s$.

When the "RECOVER" message appeared, an audio message saying "*Recover, recover*" was concurrently played inside the cockpit, ensuring that participants would start the recovery procedure immediately. During the recovery, participants were free in their handling of the aircraft. They were told to recover as they would normally do, as to not train them to perform the recovery procedure differently than their own training. It was mentioned that a consistent recovery was desired. This was one of the focus points of the training as well. Finally, participants were only able to control the elevator and throttle. The control inputs on the ailerons and rudder from the yoke and pedals were ignored in the model and controlled by the autopilot (wings-level), in order to keep the stall fully symmetric.

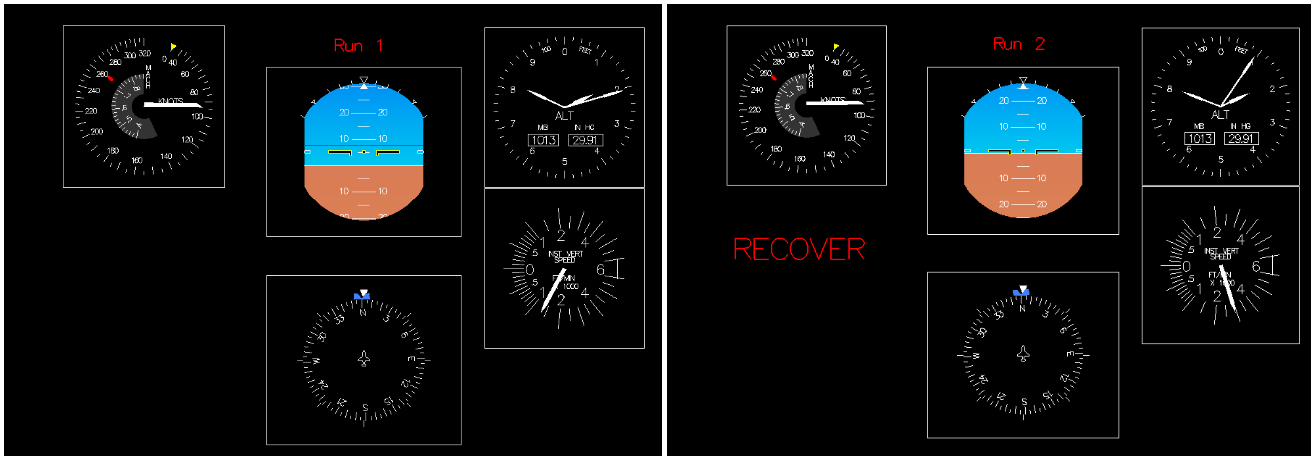
2. Passive Experiment

The passive experiment used a staircase procedure to determine the upper a_1 JND threshold. There are numerous staircase procedures [19, 27–30] that have different benefits and complexity. For this research the same method is used as in the paper by Smets et al. [13], the Parameter Estimation by Sequential Testing (PEST) procedure [19]. The PEST procedure is an adaptive staircase procedure that optimizes the stimulus level to quickly converge towards a threshold. By using this procedure, the found threshold can be compared to Smets et al. [13]. A few aspects of the procedure were carefully considered for the current experiment:

1) *When to change stimulus level* - For this a two-down, one-up (2D1U) procedure was used. This means that two correct identifications of the more abrupt stall leads to a step closer to the real parameter value, whereas one mistake leads to a step away from the real parameter value.
2) *What stimulus level to try next* - For this, the rules as laid out by Taylor and Creelman [19] were used. This means that the second step in a given direction had the same size as the first. At the fourth step in a given direction, the step size was doubled. Finally, at every reversal, the step size was halved. One exception in this is the first reversal, which is an exception also used by Smets et al. [13] and Imbrechts et al. [14]. To assist participants who

Table 4 Active experiment test order assignment from Table 3 for each participant.

Participant	Repetition				Participant	Repetition			
	1	**2**	**3**	**4**		**1**	**2**	**3**	**4**
1	A	B	H	C	9	G	D	F	E
2	B	C	A	D	10	H	E	G	F
3	C	D	B	E	11	A	F	H	G
4	D	E	C	F	12	B	G	A	H
5	E	F	D	G	13	C	H	B	A
6	F	G	E	H	14	D	A	C	B
7	G	H	F	A	15	E	B	D	C
8	H	A	G	B	16	F	C	E	D



(a) Primary Flight Display with flight director.

(b) "RECOVER" message.

Fig. 7 The primary flight display, showing the flight director for the active experiment, and the recover message that is shown in both the passive and active experiments.

made an early mistake from taking numerous steps to converge towards the threshold, this exception allowed them to take larger steps towards the threshold after an early mistake.

3) *When to terminate the staircase* - The procedure was stopped if one of the following three criteria was met. The procedure stopped if the step size was equal to or smaller than $1/64^{\text{th}}$ of the original step size, if the number of reversals reached 8, or if 32 total comparisons were made. The maximum of 32 comparisons was chosen because it matches the number of trials for the active experiment.

For this experiment, the starting stimulus level was $a_1 = 50$ with a step size of 7.5. By setting the step size to 7.5, the participant reached 27.5 after 3 steps towards the threshold, which is very close to the baseline a_1-value of 27.6711. This prevented participants from reaching four consecutive correct answers before the first reversal, and avoided an early doubling of the step size, hence ensuring an efficient staircase procedure.

The procedures of the passive experiment were the same as in the active experiment, except for the flight director shown in Figure 7a, which was removed for the passive experiment. The "RECOVER" message remained, as well as the audio message, to give participants a sense of when the autopilot initiated the recovery.

C. Apparatus

The experiment was performed in the SIMONA Research Simulator at the Faculty of Aerospace Engineering of Delft University of Technology, see Figure 8. SIMONA has a 6-degree-of-freedom hydraulic hexapod motion system. Participants were seated in the captain's seat, shown on the left in Figure 9, where a control column is present. Participants wore a noise canceling headset to mask any noise coming from the motion system's actuators. In the background, static engine noise played on the cockpit speakers throughout the experiment, to further mask the noise from the motion system.

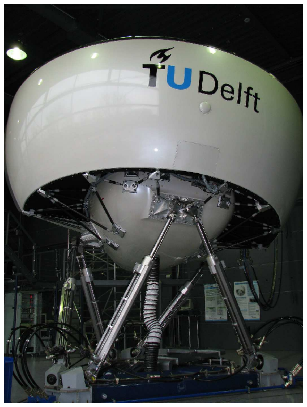


Fig. 8 The SIMONA research simulator [31].

Fig. 9 Cockpit view from inside SIMONA.

During the experiment, participants could observe the crucial flight states on the primary flight display, which showed multiple instruments as can be seen in Figure 7a. Furthermore, they had a second display that showed the engine parameters as shown in Figure 10. Furthermore, participants were provided with an outside visual covering 180° × 40°, which is provided by three projectors that have a resolution of 1280 × 1024 pixels and a refresh rate of 60 Hz [31]. The outside visual image was provided by a FlightGear visual database, set to an initial position above Amsterdam Schiphol airport.

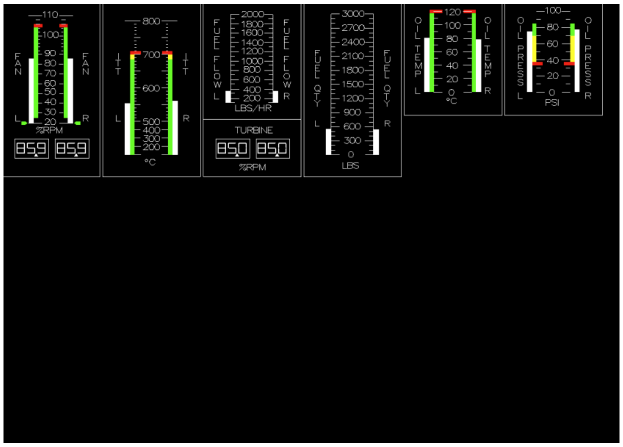


Fig. 10 Engine display as used in the experiment [14].

The settings for the washout filter used to provide simulator motion cues in the experiment are found in Table 5. These settings are equivalent to those used by Smets et al. [13] and Imbrechts et al. [14], except for the pitch gain which was set to 0.7 instead of 0.5 in the work by Imbrechts et al. [14]. The roll, yaw, and sway motions were not used for this experiment, since only symmetrical stalls were simulated.

Table 5 Experiment motion filter settings (pitch, forward, vertical).

High-pass filters						Low-pass filters	
ω_{n_q}	1.0 rad/s	ω_{n_z}	2.0 rad/s	ω_{n_x}	1.2 rad/s	ω_{n_x}	2.4 rad/s
ζ_q	0.7	ζ_z	0.7	ζ_x	0.7	ζ_x	0.7
ω_{b_q}	0.0 rad/s	ω_{b_z}	0.3 rad/s	ω_{b_x}	0.0 rad/s		
K_q	0.7	K_z	0.5	K_x	0.5		

D. Participants

A total of 16 participants participated in the experiment, all of whom had experience flying twin-turbine aircraft. The participants had a mean age of 48.5 years (standard deviation (SD) = 9.1 years), and a mean number of flight hours of $9,567$ hours (SD = $6,961$ hours). The group consisted of 10 Captains, 4 First Officers, and 2 Second Officers. Four participants had a Cessna Citation II type rating, eight had a Boeing 777/787 type rating, four had a Boeing 737 type rating, one participant had an Embraer E175, E190, and E195-E2 type rating, one participant had a Boeing 747 rating, and one participant had a Gulfstream G650 type rating. There were several participants who had multiple ratings.

Before the experiment started, it was indicated to the participants that their stall recovery performance was not evaluated and they were asked to focus on detecting differences in the model's response. Pilots voluntarily participated in this experiment and gave informed consent before starting. This research was approved by the Human Research Ethics Committee of TU Delft under application number 3643.

E. Data Analysis

1. Passive experiment

For the passive experiment, the final JND threshold of a participant was determined by averaging the staircase's a_1 values across the last three reversals, as was done by Smets et al. [13]. The resulting threshold is denoted as a_1^+, as it is the upper threshold for a_1. From the individual thresholds, the average threshold of the entire group was determined. Furthermore, the individual thresholds of each participant from Smets et al. [13] were compared to the data set of this experiment through the Mann-Whitney U test, since neither data set is normally distributed, to see if the results are significantly different.

The a_1^+ threshold that was found in this experiment is the 70.7% threshold because of the 2U1D staircase used [28], as opposed to a 50% threshold for a one-up, one-down procedure used by Smets et al. [13]. This 70.7% threshold is valid if the psychometric curve for percentage of correctness is from 0% to 100%. However, due to 2AFC design, this curve is from 50% to 100%, since the default percentage of a correct response for a two-choice experiment is 50%. Consequently, the 70.7% threshold that resulted from the 2U1D staircase is shifted to the 85.35% threshold point on the 2AFC psychometric function.

Next to this, there is a second method to determine the threshold of the participant group, which is less influenced by outliers than direct averaging. This was done by fitting the psychometric function through the average percentage of correct responses of each individual a_1-value tested in the experiment based on the combined answers of all participants. This method was also used by Smets et al. [13] and the resulting μ of both CDFs can therefore be directly compared. Because this experiment used the 2AFC question, the μ for this experiment represents 75% threshold, whereas the μ of Smets et al. [13] represents the 50% point on the CDF. However, both are the same point on the psychometric function and can therefore be compared.

2. Active experiment

The method of constant stimuli used for the active experiment, which does not give a direct threshold in its current set up [20]. Therefore, only the combined percentages of correct responses per participant for each of the eight test conditions were taken as the data set and used to estimate the JND threshold of the participant group. This was done by fitting the CDF through the average of correct responses, which resulted in the psychometric function for the active experiment. Then, the 75% threshold could also be determined by setting $P(\varphi) = 0.75$ in Equation (5) to estimate the corresponding active JND threshold.

To support this found threshold, a pairwise comparison between the data sets of each a_1 condition was done. First, the data was tested for normality with the Shapiro-Wilk test. Following this, a Friedmann ANOVA was used to determine if an overall statistically significant difference existed across all test conditions. If so, pairwise comparisons (Wilcoxon signed rank tests) between each a_1 condition and the baseline comparison case were used to highlight the active experiment conditions with a chance of detection of the elevated a_1-value that statistically exceeds the 50/50 chance level.

IV. Results

A. Representative individual pilot data

Figure 11 and 12 show representative examples of individual participant results – for Subjects 14 and 15, respectively – for both the passive (subfigure (a)) and active scenarios (subfigures (b) and (c)). Figure 11(a) and 12(a) show the full staircase completed by both participants, together with the upper JND threshold for a_1 previously measured in [13] (blue dashed line) and the current participant's JND threshold (red dashed line). The true baseline value of a_1 is indicated with the black dashed line. While Subject 14 required more trials for the staircase to converge, both participants are seen to converge to similar passive JND thresholds – i.e., 27.91 and 28.84 for Subject 14 and 15, respectively – that are both clearly lower than those reported by Smets et al. [13].

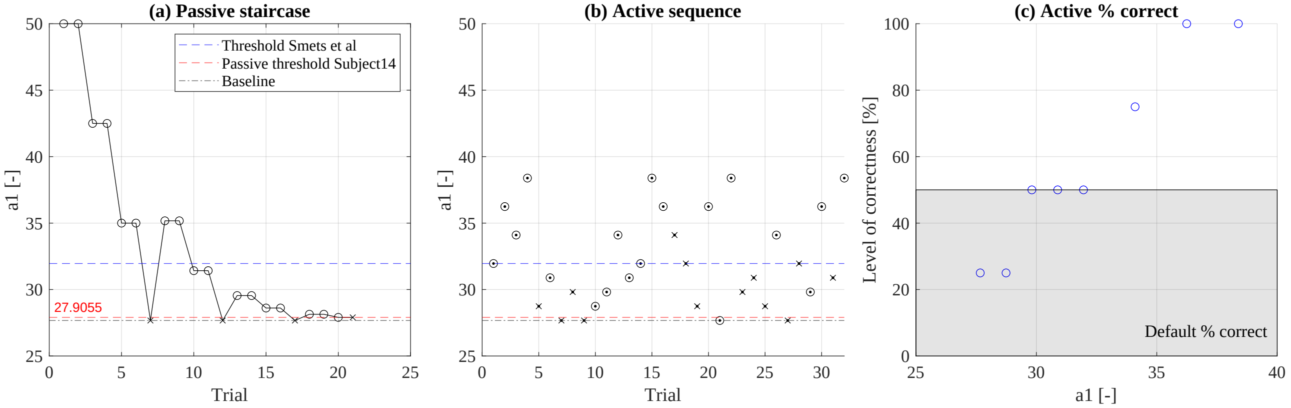


Fig. 11 Example results of both the passive (a) and active experiment (b)-(c) for Subject 14.

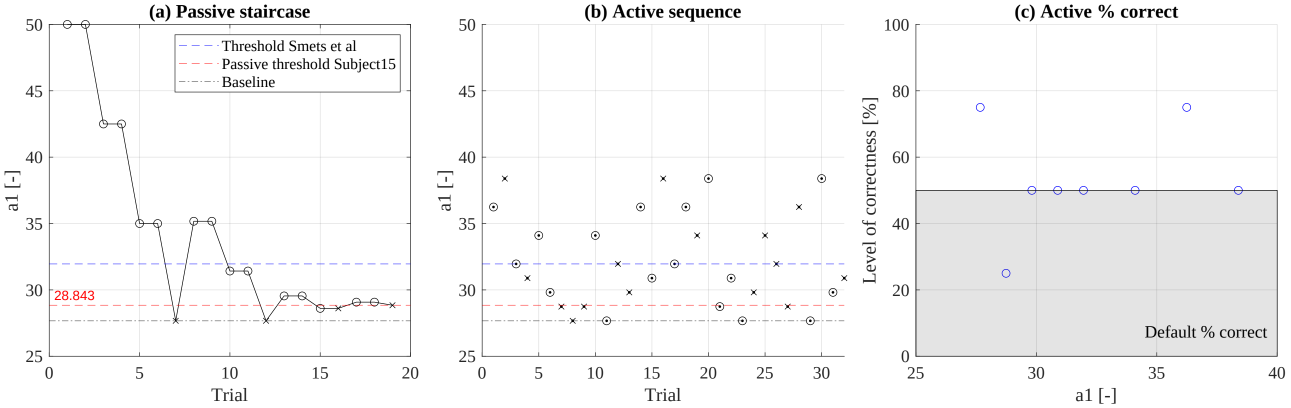


Fig. 12 Example results of both the passive (a) and active experiment (b)-(c) for Subject 15.

For the active experiment, Figure 11(b) and 12(b) show the raw outcomes of the repeated evaluation of the eight experiment conditions from Table 2 for both participants. Again, the same reference a_1 (threshold) values are shown as in the passive experiment plots. Furthermore, correctly identified higher a_1 occurrences are indicated with circular markers; incorrect baseline a_1 selections are indicated with crosses. Figure 11(c) and 12(c) show the resulting level of correctness as a function of the tested a_1 value, which summarizes the percentage of correct selections (out of four) for each test condition. The gray shaded area indicates the range of correctness percentages that indicates participants could not could not reliably identify the highest a_1 conditions, i.e., a 50/50 guessing probability and below. Subject 14's active experiment results (Figure 11(c)) show an increasing level of correctness with increasing a_1 and a correctness level well above 50% for the highest a_1 settings, as expected. For Subject 15, Figure 11(c) shows that this participant was unable to correctly identify even the highest tested a_1 cases, which may be explained by less consistent flying of the stall maneuver, see Section IV.C.2. In the following, the results of the passive and active experiment across all participants are discussed separately.

B. Passive experiment

As explained in Section III.B.2, the staircase used in the passive experiment yielded the 70.7% JND upper threshold for a_1. The measured individual thresholds were based on the average a_1 value corresponding to the last three reversals for each participant, see Figure 11(a) and 12(a). The resulting individual thresholds for all participants are shown, together with the corresponding results from [13], as boxplots in Figure 13.

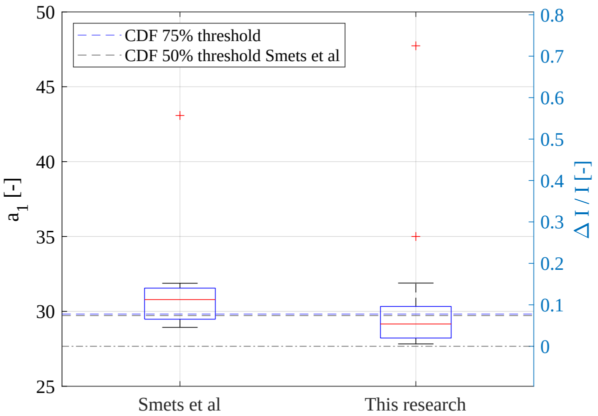


Fig. 13 Boxplot representations of the passive staircase results as well as the CDF thresholds from the work by Smets et al [13] and this research.

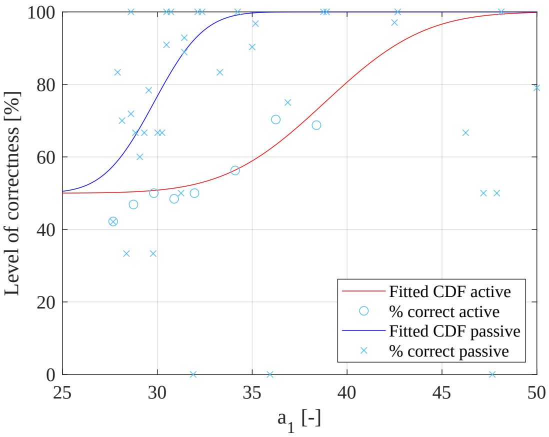


Fig. 14 CDF of the active and passive experiment, fitted through the percentage correct for each tested a_1 value.

The horizontal dashed lines in Figure 13 also indicate the JND thresholds estimated by fitting the psychometric function, as defined by its CDF in Equation (5), to the aggregated staircase data of all participants. For the current experiment, the blue CDF in Figure 14 was obtained following this methodology. As can be seen in Figure 13, the average 75% threshold obtained from the CDF for the current experiment ($P(75\%) = 29.82$ or $a_1^+ = 0.078$ when expressed as a Weber fraction) was equivalent to the 50% correctness threshold estimated in [13] ($P(50\%) = 29.72$, or $a_1^+ = 0.074$). The inconsistency with the average result obtained from the individual JND thresholds (boxplots in Figure 13) is expected and attributed to the calculation of the threshold values from the staircase data, which due to the different staircase was done differently for the current experiment and [13].

One participant (incorrectly) changed his strategy for detecting the more abrupt flow separation cases after the training phase of the passive experiment. This resulted in an erroneous first portion of this participant's staircase, as multiple wrong answers quickly resulted in the stimulus level remaining at $a_1 = 50$, the maximum allowed value, for 8 consecutive trials. After intervention by the experimenter, this participant continued the staircase as normal. For the passive experiment result in Figure 14 these 8 consecutive wrong answers at $a_1 = 50$ were omitted from the data used to fit the CDF. However, omitting this erroneous data had little influence on the outcome: with the erroneous data included the CDF mean (μ) and standard deviation (σ) are 29.8187, and 2.0772, respectively, as opposed to 29.8188 and 2.0769 with this data excluded. Overall, this shows the robustness of the JND thresholds estimated from the fitted CDFs.

C. Active experiment

1. JND threshold and statistical analysis

For estimating a JND threshold from the active experiment, the percentage of correctness for each condition of every participant was calculated as shown in Figure 11(c) and 12(c). The average percentage of correctness across all participants is shown with the circular markers in Figure 14, together with the psychometric curve fit (red line in Figure 14) to the active experiment's data. Evidently, the active experiment's CDF estimate is shifted far to the right compared to the passive experiment's CDF. From the psychometric function, the active 75% JND threshold was estimated as $P(75\%) = 38.83$ (or $a_1^+ = 0.40$ when expressed as a Weber fraction). This is markedly higher than the passive JND threshold of $P(75\%) = 29.82$ (or $a_1^+ = 0.078$), which is in line with Hypothesis **H1**.

To further quantify the statistical reliability of the active experiment data, the percentage of correctness for each condition for each participant (i.e., 16 values) was taken as a sample for statistical analysis. First, all samples were tested for normality using the Shapiro-Wilk test, from which only the data for conditions 3 and 5 (+0.5 JND and +1 JND, respectively) was found to be normally distributed. Hence, a Friedmann ANOVA was used to test for an overall significant effect across all experiment conditions. The Friedman test indeed confirmed a statistically significant effect of the a_1 setting on the correctness levels, $\chi^2(7) = 16.105$, $p = 0.024$.

Following this, pairwise comparisons between the different conditions were performed using Wilcoxon signed-rank tests, see Table 6. This analysis showed that no significant differences exist between the level of correctness measured for Conditions 1-6. However, Condition 7 shows a significant difference with Conditions 1-5, while Condition 8 shows a significant difference with Conditions 1 and 5, as well as overall lower p-values than found for Conditions 1-6. Overall, this indicates that in the active experiment the participants only reached a level of correctness significantly higher than the 50% chance level for Conditions 7 and 8, corresponding to very high a_1 settings of +2.0 JND and +2.5 JND, respectively (see Table 2). This result is fully consistent with the fitted CDF for the active scenario in Figure 14.

Table 6 Pairwise comparison p-values from Wilcoxon signed-rank tests; statistically significant differences highlighted in bold.

	2	3	4	5	6	7	8
1	0.559	0.257	0.113	0.516	0.073	**0.005**	**0.017**
2		0.775	0.677	0.981	0.441	**0.030**	0.078
3			1.000	0.807	0.636	**0.020**	0.156
4				0.844	0.613	**0.010**	0.063
5					0.560	**0.012**	**0.022**
6						0.053	0.128
7							0.766

2. Control consistency analysis

To verify the potential impact of variations in the flown stall maneuver in the active experiment on its outcomes (active JND threshold), a quantitative analysis of each participants' control consistency was performed. This was done by calculating the average control input for every trial with the baseline a_1 setting at each time sample. Then, the root mean square (RMS) difference between the data of each individual trial and this average was calculated. The average RMS across all trial for each participant is taken as a metric for each participant's control consistency. Figure 15 and 16 show the control inputs δ_e, pitch attitude θ, and flow separation state X for all trials performed by Subject 14 and 15, respectively. In line with the corresponding example active experiment outcomes in Figure 11(c) and 12(c), Subject 14 shows markedly more consistent (RMS = 0.015 [rad]) control behavior than Subject 15 (RMS = 0.027 [rad]). The results for all participants are summarized in Table 7. While in some cases an inverse correlation seems to be present between the RMS magnitude and the percentage of correctness, this is certainly not consistent across all participants (e.g., Subject 12, who showed very consistent control behavior, but still very low percentages fo correctness for Conditions 7 and 8.

Table 7 Active experiment control behavior consistency in terms of the RMS of the elevator input variation and level of correctness for Condition 7 and 8 for each participant.

Participant	**RMS** $[rad]$	**% correct C7**	**% correct C8**
1	0.0279	75	25
2	0.0335	50	75
3	0.0211	75	75
4	0.0224	75	100
5	0.0228	75	100
6	0.0187	100	100
7	0.0188	100	75
8	0.0367	75	75
9	0.0187	75	75
10	0.0212	50	25
11	0.0172	50	50
12	0.0177	25	25
13	0.0192	75	100
14	0.0154	100	100
15	0.0265	75	50
16	0.0228	50	50

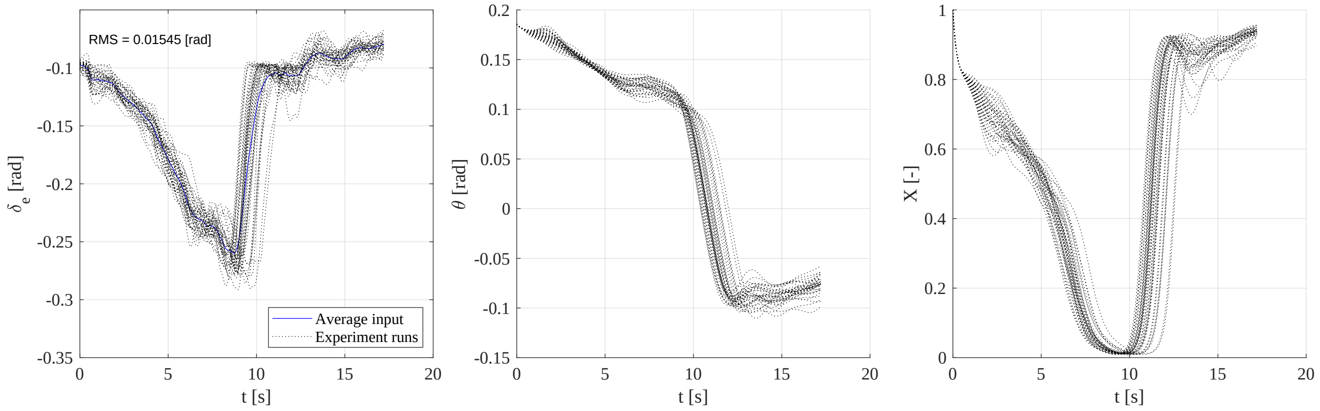


Fig. 15 Active experiment elevator input, pitch angle, and flow separation state time traces for all baseline runs from Subject 14.

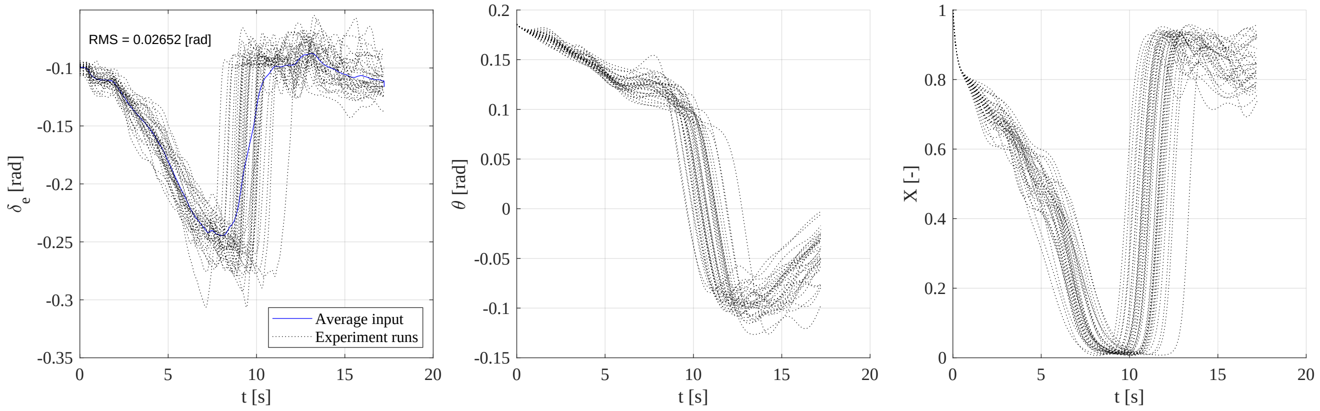


Fig. 16 Active experiment elevator input, pitch angle, and flow separation state time traces for all baseline runs from Subject 15.

D. Summary of debriefing participant comments

Apart from the data analysis, the comments participants gave after completing the experiment were aggregated and analyzed. All participants unanimously indicated that they had more difficulty to detect the differences between the compared a_1 settings in the active experiment. They explained that they felt that they had less mental capacity to focus on detecting the differences, as they were busy controlling the aircraft. Furthermore, they had the idea that their own control inputs influenced the sensation of stall abruptness, which caused them to doubt whether or not the difference they felt was due to their inputs or due to the differences in the model. Finally, after learning the true intent and potential implications of the experiment, the participants agreed that despite the considerable variation in tested a_1-values, all of the conditions presented to them in the active experiment could be used for stall training.

V. Discussion

The main objective of the experiment described in this paper was to quantify the difference between pilots' Just Noticeable Difference (JND) thresholds for an aircraft model's stall abruptness parameter (a_1) between passive and active stall scenarios. For this, an experiment with 16 active pilots was performed, whose sensitivity for detecting adjusted a_1 settings was measured using well-known psychophysical paradigms in both passive (observer) and active (pilot-in-command) settings.

Based on earlier experiments into human motion perception thresholds [16, 17], it was expected that higher JND thresholds for stall abruptness would be found in an active scenario than in a passive scenario. The obtained experiment data indeed show that the entire psychometric function of the active scenario is shifted well to the right of the passive psychometric function (Figure 14). The 75% thresholds of both CDFs lie at $P(75\%) = 29.82$ and $P(75\%) = 38.83$ for the passive and active cases, respectively; or 0.078 and 0.40 when expressed as Weber fractions. Overall, this indicates that the JND threshold for the active experiment is over five times higher, which shows that the sensitivity of participants for changes in stall abruptness for pilots flying a stall themselves is greatly decreased. Consequently, Hypothesis **H1** is accepted. This result implies that engagement in active control can 'mask' deficiencies in the model that is controlled and hence that aircraft models used for active flight seem to allow for considerable parameter offsets – larger even than previously reported passive JND values [13, 15] – before these are consistently noticed by pilots.

The passive part of the current experiment was a (partial) replication of the earlier investigation by [13] and therefore it was hypothesized that the results (average JND) would be equivalent. However, the current experiment implemented an improved staircase method for the passive scenario compared to earlier investigations [13, 14]. Consequently, only the psychometric function CDF that is fitted through the aggregated data of all participants allows for a direct comparison of the results. The average 75% threshold obtained from the CDF for the current experiment ($P(75\%) = 29.82$ or $a_1^+ = 0.078$ when expressed as a Weber fraction) was equivalent to the 50% correctness threshold estimated in [13] ($P(50\%) = 29.72$, or $a_1^+ = 0.074$). Therefore, hypothesis **H2** is accepted.

When comparing the resulting individual JND thresholds obtained directly from each participant's staircase data, a higher threshold value was expected for the current experiment, as the improved staircase procedure estimates the 70.7% correctness threshold, whereas Smets et al. [13] measured the 50% correctness threshold. However, on average the individual thresholds measured in this study ($a_1^+ = 0.11 \pm 0.094$) were in fact found to be lower than those reported in Smets et al. ($a_1^+ = 0.16 \pm 0.14$). While this difference was not found to be statistically significant (Mann-Whitney U test, $p = 0.076$), there are different possible explanations for this result. Firstly, Smets et al. asked their participants if they noticed a difference between two test runs. This question formulation for the pairwise comparisons left more room for interpretation for their participants (i.e., there could be differences possible in all aspects of the simulation) compared to asking participants to identify the run with the most abrupt flow separation, as done in the current experiment. The 'Yes/No' question of Smets et al. is particularly prone to biases, as participants may be more inclined to answering 'Yes', although they did not feel any difference or were unsure if there was a difference to be felt [26]. Smets et al.'s solution was to implement a comparison between two baseline models for every third trial in the staircase, to minimize participants' bias in indicating that they did feel a difference. However, as a result, a third of their measurements was dedicated to detecting the bias, whereas the current methodology used every measurement to converge towards the threshold, which may have caused some confusion and less reliable staircase procedure outcomes. Next to these effects of the different staircase, also the stall model was slightly modified compared to [13] (addition of C_{m_q}) and the use of a different sample of pilots may also have contributed to the difference in passive experiment outcomes.

Using the method of constant stimuli only for the active scenario in the current experiment may, in hindsight, have been a confusing choice for the participants, which may directly contribute to the higher active JND thresholds found. One of the reasons is that, since participants randomly started with one of the Latin squares with active conditions, there was no direct comparison from the training to the real experiment. In the training, participants trained with detecting differences between the most extreme a_1 values they would encounter, which are $a_1 = 50$ $[-]$ and $a_1 = 38.3869$ $[-]$ for the passive and active experiment, respectively. For the passive experiment, the first runs also contained the comparison with $a_1 = 50$ $[-]$, so they could directly recognize the cues from the training. For the active experiment, this was not necessarily true. For instance Subject 1 and 11 and Subject 8 and 10 started with Latin square A and H, respectively. Both Latin squares started with relatively low a_1 values in the first three or four runs. These subjects did, therefore, not recognize any of the cues that they experienced during the training, which can be solved by using a staircase procedure.

This confusion became evident from the responses of the participants during the debriefing. Most participants indicated that they felt that they guessed for most of the experiment and were unsure about their performance for the active part of the experiment. Another contributing factor is that the highest a_1 value used in the experiment is $a_1 = 38.3869$, whereas the resulting 75% threshold of the combined answers is $P(75\%) = 38.83$. The used conditions were, therefore, insufficient to assess the entire psychometric function for the active task, since the entire CDF curve is based on half of the threshold information. Therefore, for future research, it is better to assess the threshold for active flying including also higher values of a_1.

Unfortunately, due to the fact that the highest a_1 condition was not sufficient to reach the 100% correctness level for the active experiment, the analysis of the control behavior cannot lead to conclusive answer on what the influence of control behavior variations of the participants is on the resulting percentages correct. Furthermore, because only 4 repeated measurements were taken for each participant and a_1 condition, the individual level of correctness percentages are a very course representation of the true underlying probability and very susceptible to effects of control (in)consistency. However, it is an interesting aspect to investigate in future research. For instance, when an individual threshold is found, the consistency in control behavior can be directly compared to verify whether or not there is an influence on the obtained thresholds.

One of the points of feedback raised by the participants regarding the simulated scenario was its odd entry into stall. The simulation from Smets et al. [13], which was used in this research as well, started at 10° pitch up and lowered the pitch attitude to about 5° to reach a 1 kts/s deceleration into the stall. Participants mentioned that their usual stall training includes a pitch-up behavior due to, for instance, the altitude hold mode of the autopilot, a scenario used by Stepanyan et al. [32] and Lombaerts et al. [33] in their stall simulation experiments. This subtle change to how the stall is induced would create a more realistic scenario for the participants.

In the debriefing with the participants, the cues on which the participants based their answer were discussed. They indicated that the main cues came from the primary flight display, as well as the stall buffet. No participant indicated that they used the pitch motion feedback, which was accidentally slightly different (higher gain) compared to [13, 14]. It is, therefore, assumed that this difference had no significant influence on the outcomes reported in this paper.

Finally, a shortcoming of the active experiment was that no force feedback was implemented on the column used for providing elevator inputs. The control column in SIMONA was configured as a passive a mass-spring damper system. As a result, participants could not feel the elevator feedback as they would normally have in the real Cessna Citation II. Furthermore, all participants fly the Cessna Citation II, Boeing 737, 777/787, or the Embraer family, all of which have a control column with force feedback. In future research, implementing force feedback may contribute to a more complete feel of the aircraft. Some of the participants indicated that they already noticed a difference in the amount of back-pressure they needed to apply between the test runs. Hence, especially JND thresholds for parameters as the a_1 abruptness parameter considered in this paper may be affected by the presence of force feedback, so adding this to the simulation is an essential step for future experiments.

This paper focused fully on measuring the upper JND threshold for a single stall model parameter, i.e., the a_1 stall abruptness parameter of the model derived in [18]. It is essential that similar experimental data regarding pilots' sensitivity to changes in model parameter values is also expanded to other crucial stall parameters – such as time constants that model lags in flow separation and hysteresis, e.g., τ_1 in [13]. Furthermore, as pilots may be more sensitive to a decrease than to an increase in any model parameter (or vice versa), it is also important that both the 'upper' and 'lower' JND thresholds are measured. Finally, to verify to what extent measured JND thresholds – as well as the relative offset of passive and active thresholds – vary as a function of aircraft type and size, a direct comparison experiment with models representing different aircraft is a crucial next step.

Finally, the found sensitivity of the participants can be related to the uncertainty of the model as created by Van Ingen et al. [18]. They found the baseline value for $a_1 = 27.6711$ with standard deviation $\sigma = 6.72$ (or 0.248 when expressed as a fraction of the mean). Brill et al. [23] used a slice-based modeling approach to identify the a_1 value for the Cessna Citation II aircraft and found a value of $a_1 = 34.1856$, which is a difference of 23.5% as compared to the value used in this research. Both these uncertainties lie within the found active JND threshold of $a_1^+ = 0.40$ found by this research. This implies that the modeling techniques as used by Van Ingen et al. [18] and Brill et al. [23] can provide a stall model that is, for pilots controlling it, indistinguishable from the real aircraft and would, hence, be able to provide effective simulator-based training.

The results of this research show that pilots' sensitivity to changes in simulated stall abruptness strongly decreases when they are in active control of the stall model, compared to a passive observer setting. Furthermore, the quantitative JND threshold data can contribute to the further development of truly quantitative requirements on the required accuracy for stall models to be used in (training) simulators. Hopefully, this will enable transforming statements such as the "*within the realms of confidence*" [9] as currently mentioned included in [9] into an actual required numerical accuracy, which is crucial for ensuring simulator-based stall training is truly effective, but also not more costly than it needs to be.

VI. Conclusion

This research provides, for the first time, a quantitative estimate on how the upper Just Noticeable Difference (JND) thresholds for changes in stall abruptness in a passive observer scenario translate to an active pilot-in-command setting. For this, an experiment in TU Delft's SIMONA Research Simulator was performed with 16 active pilots, who experienced and flew symmetric stall scenarios simulated with a Cessna Citation II stall model. Across the data from all participants, the passive JND threshold was found to be 29.82, or 0.078 expressed as Weber fraction compared to the baseline parameter value of 27.67. This passive JND threshold is marginally lower than the value reported in previous research, most likely due to the use of an improved staircase procedure for the passive experiment. Furthermore, this research successfully measured an active JND threshold for stall abruptness, which lies at 38.83, or 0.40 expressed as a Weber fraction. Hence, the active threshold is found to be over five times higher than the passive threshold, indicating that pilots have a strongly decreased sensitivity for changes in stall abruptness when flying a stall themselves. The obtained active JND thresholds provide a basis for regulatory bodies to define and implement more precise accuracy standards for stall model fidelity in simulators, to further optimize the effectiveness of simulator-based stall training.

Acknowledgments

The authors would like to thank the technicians and responsible personnel of the SIMONA Research Simulator at TU Delft for their help performing the experiments, as well as all pilots who volunteered as participants for this study. Furthermore, the first author would like to thank the AIAA's Modeling and Simulation Technologies TC for awarding him the AIAA Luis de Florez Graduate Award for the research project that has resulted in this paper.

References

[1] IATA, "Loss of Control In-Flight Accident Analysis Report," Tech. rep., International Air Transport Association, 2019.

[2] "ICAO Doc 10011: Manual on Aeroplane Upset Prevention and Recovery Training," Tech. rep., International Civil Aviation Organization, 2014.

[3] NTSB, "Aircraft Accident Report; Loss of Control on Approach; Colgan Air, Inc," Tech. rep., National Transportation Safety Board, Washington, DC, 2 2010.

[4] OVV, "Turkish Airlines Flight TK1951, a Boeing 737-800, Crashed during approach," Tech. rep., Dutch Safety Board, 2010. URL `www.safetyboard.nl`.

[5] d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, B., "Final Report on AF447," Tech. rep., Ministère de l'Écologie, du Développement durable, des Transports et du Logement, 2009. URL `www.bea.aero`.

[6] "Loss of control prevention and recovery training," `https://www.easa.europa.eu/en/downloads/71682/en`, feb 2019.

[7] "Upset Prevention and Recovery Training," `https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-111_CHG_1.pdf`, jan 2017.

[8] "Part 60 - Flight Simulation Training Device Initial and Continuing Qualification and Use," Tech. rep., Federal Aviation Administration, United States Department of Transportation, 10 2021. URL `https://www.ecfr.gov/current/title-14/part-60`.

[9] EASA, "Certification Specifications for Aeroplane Flight Simulation Training Devices 'CS-FSTD(A)'," Tech. rep., European Aviation Safety Agency, 2018.

[10] J. A. Schroeder and J. Bürki-Cohen and D. A. Shikany and D. R. Gingras and P. Desrochers, "An Evaluation of Several Stall Models for Commercial Transport Training," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, National Harbor (MD)*, 2014. https://doi.org/10.2514/6.2014-1002.

[11] Grant, P. R., Moszczynski, G. J., and Schroeder, J. A., "Post-stall flight model fidelity effects on full stall recovery training," *Proceedings of the AIAA Modeling and Simulation Technologies Conference*, 2018. https://doi.org/10.2514/6.2018-2937.

[12] Cunningham, K., Shah, G. H., Murphy, P. C., Hill, M. A., and Pickering, B. P., "Pilot sensitivity to simulator flight dynamics model formulation for stall training," *Proceedings of the AIAA Scitech 2019 Forum*, 2019. https://doi.org/10.2514/6.2019-0717.

[13] Smets, S., De Visser, C. C., and Pool, D. M., "Subjective Noticeability of Variations in Quasi-Steady Aerodynamic Stall Dynamics," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, San Diego (CA)*, 2019. https://doi.org/10.2514/6.2019-1485.

[14] Imbrechts, A., De Visser, C. C., and Pool, D. M., "Just Noticeable Differences for Variations in Quasi-Steady Stall Buffet Model Parameters," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, San Diego (CA)*, 2022. https://doi.org/10.2514/6.2022-0510.

[15] De Visser, C. C., and Pool, D. M., "Stalls and Splines: Current Trends in Flight Testing and Aerodynamic Model Identification," *Journal of Aircraft*, Vol. 60, No. 5, 2023, pp. 1480–1502.

[16] Hosman, R. J. A. W., and van der Vaart, J. C., "Vestibular Models and Thresholds of Motion Perception. Results of tests in a flight simulator," Tech. rep., Delft University of Technology, Delft, 1978. URL http://resolver.tudelft.nl/uuid:72bb1e55-7304-459f-a47c-dae7984418e3, lR-265.

[17] Valente Pais, A. R., Pool, D. M., De Vroome, A. M., Van Paassen, M. M., and Mulder, M., "Pitch motion perception thresholds during passive and active tasks," *Journal of Guidance, Control, and Dynamics*, Vol. 35, No. 3, 2012, pp. 904–918. https://doi.org/10.2514/1.54987.

[18] Van Ingen, J., Visser, C. C., and Pool, D. M., "Stall Model Identification of a Cessna Citation II from Flight Test Data Using Orthogonal Model Structure Selection," *Proceedings of the AIAA Scitech 2021 Forum*, 2021. https://doi.org/10.2514/6.2021-1725.

[19] Taylor, M. M., and Creelman, C. D., "PEST: Efficient estimates on probability functions," *The Journal of the Acoustical Society of America*, Vol. 41, No. 4A, 1967, pp. 782–787.

[20] Simpson, W. A., "The method of constant stimuli is efficient," *Perception & psychophysics*, Vol. 44, 1988, pp. 433–436.

[21] Van Horssen, L., De Visser, C. C., and Pool, D. M., "Aerodynamic Stall and Buffet Modeling for the Cessna Citation II Based on Flight Test Data," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, Kissimmee (FL)*, 2018. https://doi.org/10.2514/6.2018-1167.

[22] Fischenberg, D., "Identification of an Unsteady Aerodynamic Stall Model from Flight Test Data," *Proceedings of the AIAA Atmospheric Flight Mechanics Conference, Baltimore (MD)*, 1995, pp. 138–146. https://doi.org/10.2514/6.1995-3438.

[23] Brill, P. A. R., Pool, D. M., and de Visser, C. C., "Improved Kirchhoff Stall Model Parameter Estimation Accuracy through Optimal Data Slicing," *Proceedings of the AIAA Atmospheric Flight Mechanics Conference, Orlando (FL)*, 2025.

[24] Mulder, J., van Staveren, W., van der Vaart, J., de Weerdt, E., de Visser, C., in 't Veld, A., and Mooij, E., "Flight Dynamics," Faculty of Aerospace Engineering, Delft University of Technology, 3 2013. Lecture Notes AE3212.

[25] Van den Hoek, M., de Visser, C., and Pool, D., "Identification of a Cessna Citation II Model Based on Flight Test Data," *Fourth CEAS Specialist Conference on Guidance, Navigation and Control*, Springer, 2017, pp. 259–277.

[26] Kingdom, F. A., and Prins, N., *Psychophysics: A practical introduction*, 2$^{\mathrm{nd}}$ ed., Mica Haley, 2016.

[27] Cornsweet, T. N., "The Staircase-Method in Psychophysics," *The American Journal of Psychology*, Vol. 75, No. 3, 1962, pp. 485–491. URL http://www.jstor.org/stable/1419876.

[28] Levitt, H., "Transformed up-down methods in psychoacoustics," *The Journal of the Acoustical society of America*, Vol. 49, No. 2B, 1971, pp. 467–477.

[29] Leek, M. R., "Adaptive procedures in psychophysical research," *Perception & Psychophysics*, Vol. 63, No. 8, 2001, pp. 1279–1292.

[30] Pentland, A., "Maximum likelihood estimation: The best PEST," *Perception & Psychophysics*, Vol. 28, No. 4, 1980, pp. 377–379.

[31] Stroosma, O., van Paassen, M. M., and Mulder, M., "Using the SIMONA Research Simulator for Human-machine Interaction Research," *Proceedings of the 2003 AIAA Modeling and Simulation Technologies Conference and Exhibit*, 2003. https://doi.org/10.2514/6.2003-5525.

[32] Stepanyan, V., Krishnakumar, K. S., Kaneshige, J., and Acosta, D. M., "Stall Recovery Guidance Algorithms Based on Constrained Control Approaches," *Proceedings of the AIAA Guidance, Navigation, and Control Conference, San Diego (CA)*, 2016. https://doi.org/10.2514/6.2016-0878.

[33] Lombaerts, T., Schuet, S., Stepanyan, V., Kaneshige, J., Hardy, G., Shish, K., and Robinson, P., "Piloted simulator evaluation results of flight physics based stall recovery guidance," *Proceedings of the AIAA Guidance, Navigation, and Control Conference, Kissimmee (FL)*, 2018. https://doi.org/10.2514/6.2018-0383.