Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 250–261
https://doi.org/10.54941/ahfe1008259

Measuring Mental Workload in Immersive Fighter-Flight Simulation: A Multimodal Approach

Jean-Christophe Hurault, Marianne Jarry, Grégory Froger, Anne Lise Marchand, and Colin Blättler

CREA, Centre de recherche de l'École de l'Air et de l'Espace, École de l'Air et de l'Espace, F-13660 Salon-de-Provence, France

ABSTRACT

Workload indicators in fighter aviation rarely agree, and it is not known in advance which one to trust. This study examined how subjective, performance and oculometric indicators associate, dissociate or remain insensitive in immersive fighter-flight simulation. Fifty-five military pilot trainees flew a virtual-reality mission crossing three flight phases (autopilot, low-altitude terrain following, air defense) with three secondary-task levels (memorization alone, or with one or two spatial decisions every 30 s), a 3 × 3 within-subjects design. Perceived workload (NASA-TLX) rose with both factors. Near-display gaze was the most sensitive indicator of task level, separating all three levels even outside the decision windows, yet it fell across flight phases, moving opposite to perceived workload on that factor. Decision performance did not change: Bayesian analyses favored the absence of an effect for every term except task level on accuracy, where they were inconclusive, and performance was stable at a modest level, with little room to degrade. SWORD ratings and gaze both revealed effort shifting toward the decision task. Recall of a concurrent memorization task registered the demand of flying (confounded with time on task), and the added decisions only in autopilot, where capacity remained to be consumed. Pupil diameter decreased as demand increased, a reversal attributable to the light emitted by the decision task rather than to cognitive load. Each indicator thus answered a different question, and which one depended on the factor varied. No single measure suffices in a realistic cockpit; this study maps what each one answers, for which factor, and when it can be read at all.

Keywords: Mental workload, Multimodal assessment, Immersive flight simulation, Eye tracking, NASA-TLX

INTRODUCTION

Mental workload (MWL) is central to operator performance in complex, safety-critical systems, yet it remains difficult to measure, especially in fighter aviation, where pilots face concurrent perceptual, motor, memory and decision demands, and where workload is best assessed in realistic simulated environments rather than isolated laboratory tasks (Mohanavelu et al., 2020).

Received September 25, 2026; Revised October 30, 2026; Accepted November 10, 2026; Available online December 1, 2026

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

A broad consensus treats MWL as a multidimensional construct arising from the interaction between task demands and the operator's limited processing resources (Longo et al., 2022). The multiple-resource framework (Wickens, 2008) makes this explicit: performance depends on the resources invested, their efficiency, and the competition for shared resources across perceptual, cognitive and response structures. No single indicator can therefore capture MWL on its own, a point already made nearly twenty years ago by Cegarra and Chevalier (2008), and combining them is not a matter of adding sensitive indicators, because they do not always agree. Measures associate when they rise together with demand, show insensitivity when they stay flat, and dissociate when they move in opposite directions (Hancock, 2017). Dissociation is theoretically expected and not an artifact: Yeh and Wickens (1988) showed that performance and subjective ratings diverge precisely when operators invest additional resources to protect a resource-limited primary task.

Physiological indices add promise but also difficulty. The task-evoked pupillary response tracks processing and memory load under controlled conditions (Kahneman & Beatty, 1966), but in immersive settings the pupil is dominated by the light reflex and by a continuously changing scene (Mathôt & Vilotijević, 2023). In a recent multimodal study a demanding dual task produced constriction instead of dilation, and the authors equated screen illumination across tasks to protect the measure (Dhengre & Rothrock, 2025), a control hardly compatible with high-fidelity flight simulation, where the tasks themselves are luminous.

Subjective ratings answer a different question: the NASA-TLX (Hart & Staveland, 1988) reports the experience of load rather than its consequences, while SWORD (Vidulich et al., 1991) captures allocation, not intensity. Performance measures are often least sensitive to moderate changes in demand, because operators maintain them by investing effort (Yeh & Wickens, 1988). What none of them states is how much capacity remains, which is classically estimated from the concurrent task an operator gives up first (Wickens, 2008). Whether recall of a concurrent memorization task behaves as an index of the capacity that remains is examined below.

The present study examines how these indicators associate, dissociate or remain insensitive during an immersive virtual-reality fighter-flight simulation. Load was manipulated within participants along two dimensions, flight phase and secondary-task level, yielding nine conditions. The aim is not to identify a single "best" measure, nor to combine them into one score, but to characterize where they agree, where they diverge, and what that divergence implies for what can currently be measured.

METHOD

Participants

Fifty-eight military pilot trainees were recruited. Data from three of them were lost because of file corruption during recording, for reasons unrelated to participant characteristics or performance, and these participants were excluded from all analyses. The final sample comprised 55 trainees (53 men,

2 women; mean age = 23.5 years, SD = 2.1; mean flight experience = 131.2 hours, SD = 61.8), each of whom completed all nine experimental conditions. The study was approved by the ethics committee of the French Military Health Academy (Comité d'Éthique de l'Académie de Santé des Armées, École du Val-de-Grâce, IRB00014579, IORG0011755), and all participants provided written informed consent.

Apparatus, Design and Tasks

Participants flew a Mirage 2000C in the Mission Combat Simulator, displayed in a Varjo XR-4 virtual-reality headset and controlled through a Thrustmaster HOTAS (Figure 1). The mission was flown at night with night-vision goggles, to keep scene luminance as uniform as possible. The headset's eye tracker recorded pupil diameter, binocular gaze direction and vergence-based focus distance at 200 Hz. Two virtual displays, left and right in the cockpit, carried the secondary tasks, rendered at about 50 cm, in the same near-field depth range as the instrument panel. All data streams shared a common clock.

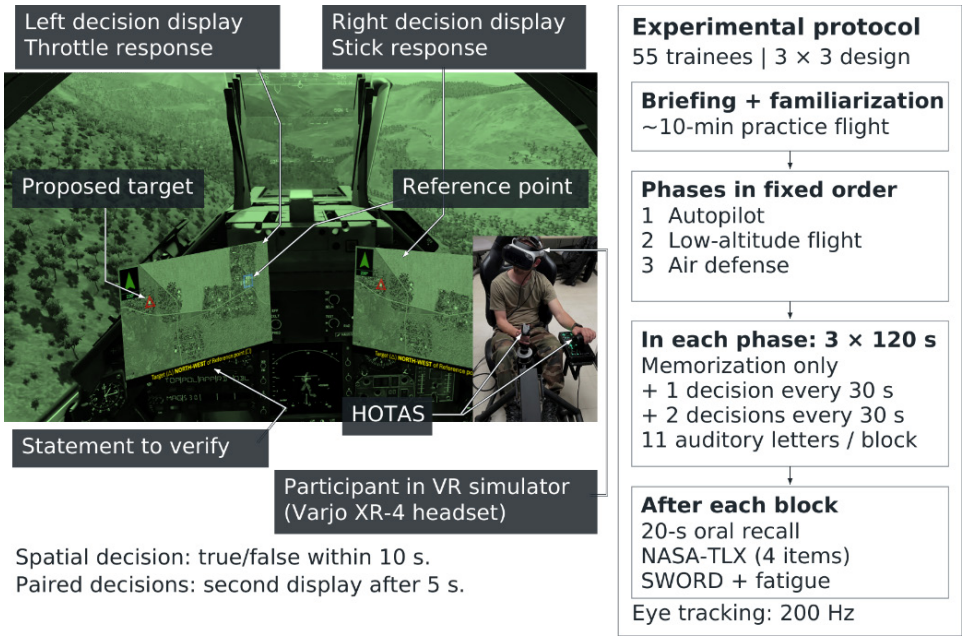


Figure 1: Simulator setup and experimental protocol. The cockpit view illustrates low-altitude flight at the two-decision task level, with paired spatial decisions presented on the left and right displays. Annotations identify the task elements and response controls; the right panel summarizes the protocol.

The three flight phases were flown in a fixed order: autopilot (cruise at 450 kt and 5,000 ft), low-altitude terrain following (below 1,000 ft above ground level), and air defense (engaging hostile aircraft to protect a ground area). The air-defense scenario was scripted, the same aircraft on the same

trajectories for all participants, although the outcome depended on the pilot. Within each phase, participants completed three 120-s blocks of increasing secondary-task level, also in a fixed order (Figure 2), after a briefing and a familiarization flight of about 10 min with practice of both tasks. The session lasted about 60 min, each block being followed by the oral recall and the questionnaires.

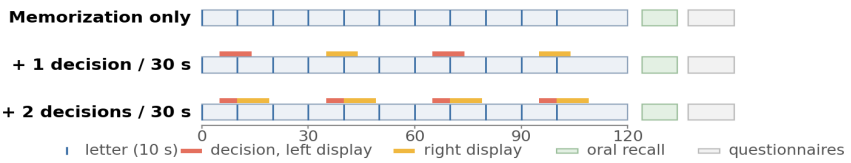


Figure 2: Structure of a 120-s block at each task level, flown in each phase. Time in seconds from the start of the block.

In all blocks, participants memorized eleven letters from the aeronautical phonetic alphabet, presented auditorily every 10 s, and recalled them orally at the end of the block. In the one-decision level, a spatial decision was added every 30 s (four per block, alternating between displays): a tone and a brief light alert on one display were followed by a map and a statement about a bearing read from it, judged true or false within 10 s with a button on the throttle (left display) or the stick (right display). True and false statements were equally frequent, so chance is .50; the map disappeared once a response was given, no feedback was provided, and a response on the other inceptor counted as an error. Distinct map items were used, none repeated within a participant, each assigned to one display. Alerts fell 5 s off the letter grid, so that at the one-decision level none coincided with a letter. In the two-decision level, decisions came in pairs, the second appearing 5 s after the first on the opposite display and therefore on the grid; since 79% of answered first decisions took longer than 5 s, the two maps were usually displayed together. The first always appeared on the left, so side, order and temporal position are confounded.

Measures

Subjective. After each block, participants completed a four-item NASA-TLX (mental demand, temporal demand, performance, effort); physical demand and frustration were omitted to keep the questionnaire short enough to repeat after every block. The workload composite is the unweighted mean of the four items, with performance reversed. SWORD weights came from the geometric means of the pairwise comparison matrix; consistency ratios were not analyzed. In memorization-only blocks only two tasks were present, so SWORD weights are compared between the one- and two-decision levels only. A single-item question on fatigue was added as a control for time on task.

Performance. Memory span was the number of letters recalled in correct serial position during a 20-s oral recall window; with eleven letters drawn from a 26-letter alphabet, chance is 0.42 of 11. The number of letters produced, whatever their position, was scored alongside it. Decision performance comprised the omission rate (no response within 10 s) and accuracy, over all decisions and over answered decisions only; both are undefined in memorization-only blocks, and accuracy on answered decisions also when every decision of a cell is omitted, which leaves 52 participants for it. Flight performance was quantified in low-altitude flight only, as the proportion of time above the 1,000-ft ceiling, for the 49 participants whose telemetry covered more than 90% of all three blocks.

Oculometric. Pupil data were preprocessed following Kret and Sjak-Shie (2019) and Mathôt and Vilotijević (2023) (artifact rejection, interpolation of blinks up to 500 ms, binocular averaging, 4-Hz low-pass filter, resampling at 100 Hz); blocks retaining less than 90% valid samples were discarded. Near-display gaze was the proportion of time the focus distance fell below 1 m, that is, the time the eyes converged inside the cockpit rather than on the outside scene. Measures were computed over each 120-s block, and for pupil diameter and near-display gaze also within and between decision windows.

Analyses

Each measure was analyzed with a repeated-measures ANOVA with flight phase (3) and secondary-task level (3, or 2 for measures undefined in memorization-only blocks) as within-participant factors, Greenhouse-Geisser correction when sphericity was violated, Holm-corrected post-hoc comparisons and partial eta squared (ηp^2) as effect size. Simple main effects are reported with uncorrected degrees of freedom. Where an effect was absent, the evidence for that absence was quantified by a Bayesian repeated-measures ANOVA and reported as BF_{01}, favoring models without the effect over otherwise identical models containing it. All analyses were computed in JASP 0.98.1, mixed-model terms being tested with likelihood-ratio tests.

Because decision accuracy and omissions are proportions over only four or eight trials per cell, and aggregate d' is undefined when a cell contains no response of one kind, decision data were also analyzed at the trial level with generalized linear mixed models (Bernoulli family, logit link) on the 1,980 trials of the 55 participants. Omissions were modeled with a random intercept for participant, an item intercept yielding a singular fit. Signal-detection parameters came from a model of the 1,574 answered trials with crossed random intercepts for participant and item and item-type as a predictor, so that its effect tests sensitivity and the intercept, criterion; d' was computed from the estimated marginal probabilities.

Relations among indicators were examined within participants: each block value was centered on its condition mean and then on the participant's mean, so that neither the manipulation nor the participant's own level remains.

RESULTS

Perceived Workload

Perceived workload increased with flight phase, $F(1.57, 85.0) = 82.3$, $p < .001$, $\eta p^2 = .60$, from 55.0 to 63.8 to 70.5 across the three phases (all pairwise $p < .001$), and with task level, $F(1.86, 100.5) = 66.6$, $p < .001$, $\eta p^2 = .55$: by 8.4 points from memorization alone to one decision ($p < .001$) and 2.5 more from one to two decisions ($p = .005$). The interaction was not significant, $F(3.61, 194.9) = 2.29$, $p = .068$. The fatigue item, a control for time on task, rose monotonically across the nine blocks, from 24.8 to 53.1 points.

Table 1: Means (SD) by flight phase and secondary-task level, 55 participants. Span = letters in correct serial position; Letters = letters recalled in any position.

Phase / Level	TLX	Letters	Span	Near Gaze %	Pupil mm
Autopilot none	48.7 (14.6)	9.07 (2.01)	5.87 (3.83)	20.7 (21.4)	3.24 (0.42)
Autopilot +1	56.4 (11.6)	8.55 (2.31)	5.16 (3.32)	40.5 (19.9)	3.12 (0.41)
Autopilot +2	59.8 (13.2)	8.09 (2.78)	4.78 (3.49)	51.5 (20.8)	2.99 (0.44)
Low alt. none	55.4 (11.2)	8.11 (2.21)	4.87 (3.39)	3.3 (13.5)	3.41 (0.43)
Low alt. +1	67.0 (10.9)	7.76 (2.39)	4.26 (2.88)	12.6 (14.2)	3.30 (0.42)
Low alt. +2	69.1 (12.1)	7.75 (2.36)	4.31 (2.96)	21.0 (14.8)	3.20 (0.43)
Air def. none	66.0 (12.9)	6.49 (2.66)	3.31 (2.62)	4.8 (12.8)	3.12 (0.42)
Air def. +1	71.9 (10.2)	7.22 (2.21)	3.31 (2.46)	13.5 (12.1)	3.04 (0.41)
Air def. +2	73.8 (11.7)	7.02 (2.76)	3.67 (2.82)	23.5 (14.8)	3.01 (0.44)

Decision Performance

Decision performance was unaffected by either factor: the omission rate remained around 20% (SD .20 to .28) in every condition (phase $F = 0.35$, task level $F = 0.97$, both ns), and neither raw accuracy nor accuracy on answered decisions varied significantly, the largest of these effects being the task-level effect on raw accuracy, $F(1, 54) = 3.57$, $p = .064$. Stability was reached at a modest level: on answered decisions accuracy ranged from .62 to .70 against a chance level of .50 ($d' = 0.83$); over all decisions, with about a fifth unanswered, it ranged from .48 to .56, against .40 for a participant answering at chance at the same omission rate.

Bayesian analyses quantify these non-effects. For the omission rate the evidence favors the absence of both factors, $BF_{01} = 15.7$ for flight phase and 4.6 for task level, and of their interaction, 5.8. For accuracy it favors the absence of the phase effect, 7.3 on raw and 11.7 on answered decisions, and of the interactions, 10.6 and 10.3. The task-level term on accuracy is the one exception: the evidence is inconclusive, 1.6 and 2.9, and what trend there is runs toward better, not worse, accuracy.

At the trial level, omissions varied with neither factor (both $p > .15$) but were more frequent on the right display, answered on the stick, than on the left (estimated marginal probabilities .23 vs. .13, $\chi^2(1) = 28.4$, $p < .001$), a difference already present at the one-decision level; side, order and item

set are confounded. On answered decisions, neither sensitivity nor criterion varied with either factor (all $p > .09$; no Bayes factor computed).

Recall and Flight Performance

Memory span decreased with flight phase, $F(1.76, 95.3) = 18.9$, $p < .001$, $\eta p^2 = .26$, from 5.3 letters in autopilot to 4.5 in low-altitude flight and 3.4 in air defense, all phases differing, and did not vary with task level overall, $F(1.98, 106.7) = 1.81$, $p = .17$, $\eta p^2 = .03$. The number of letters produced behaved the same way, phase $F(1.82, 98.3) = 22.9$, $p < .001$, $\eta p^2 = .30$, task level $F = 0.88$, $p = .41$.

The task-level effect, however, depended on the phase. On the number of letters produced the interaction was significant, $F(3.42, 184.7) = 3.20$, $p = .020$, $\eta p^2 = .06$, and simple effects locate it: recall fell as decisions were added in autopilot, from 9.07 to 8.55 and 8.09 letters, $F(2, 108) = 3.51$, $p = .033$, but not in the two demanding phases (both ns), where it was already at its lowest from the memorization-only block onward. Positional span followed the same course in autopilot, 5.87, 5.16 and 4.78 letters, without reaching significance, $F(2, 108) = 1.85$, $p = .16$, and its interaction was not significant, so the effect is carried by the number of letters recalled and not by the span.

Flight performance did not deteriorate either: in low-altitude flight, the only phase with a quantified constraint, the proportion of time above the 1,000-ft ceiling fell as task level rose, from .221 to .138, $F(1.68, 80.5) = 6.03$, $p = .006$, $\eta p^2 = .11$.

Declared Allocation and Gaze

SWORD weights showed a redistribution of effort when decision frequency doubled. The weight of the decision task increased from one to two decisions, $F(1, 54) = 12.1$, $p = .001$, $\eta p^2 = .18$, without differing across phases, $F(1.84, 99.3) = 2.08$, $p = .14$, $\eta p^2 = .04$. Since the weights sum to one, this gain came, in autopilot, where piloting weighted little (.11), from memorization (.58 to .49); in the two demanding phases from piloting, memorization stable. The piloting weight itself rose with flight phase, $F(1.96, 105.9) = 21.6$, $p < .001$, $\eta p^2 = .29$, and fell with task level, $F(1, 54) = 6.28$, $p = .015$, $\eta p^2 = .10$.

Near-display gaze was the most sensitive indicator of task level. It increased with task level, $F(1.39, 75.1) = 180.9$, $p < .001$, $\eta p^2 = .77$, separating all three levels (+9.8 points from one to two decisions, $t(54) = 10.9$), and more so in autopilot than in the two demanding phases (Table 1), phase × task level $F(2.39, 128.9) = 17.5$, $p < .001$, $\eta p^2 = .25$. Across phases it moved the other way, from 38% of the time in autopilot to 12% in low-altitude flight and 14% in air defense, $F(1.19, 64.3) = 117.3$, $p < .001$, $\eta p^2 = .68$. Saccade rate was likewise non-monotonic across phases, $F(1.94, 104.6) = 85.5$, $p < .001$, $\eta p^2 = .61$, lowest in low-altitude flight and highest in air defense, and it rose with task level, $F(1.99, 107.7) = 96.2$, $p < .001$, $\eta p^2 = .64$, though without head-rotation compensation it is not read here as a workload index.

Part of this effect is the time needed to read the maps. Restricted to the periods between decisions it still separated all three levels, $F(1.89, 101.9) = 26.0$,

$p < .001$, $\eta p^2 = .33$, about a quarter of the whole-block increase in percentage points, and more than a third in autopilot.

Pupil Diameter

Pupil diameter decreased as task level increased, $F(1.46, 78.6) = 120.0$, $p < .001$, $\eta p^2 = .69$, by 0.25 mm in autopilot, 0.22 mm in low-altitude flight and 0.11 mm in air defense, the reverse of the dilation expected from cognitive load (as in Dhengre & Rothrock, 2025). Contrasting decision windows with the periods between them located the effect: the pupil was markedly constricted during decision windows, $F(1, 53) = 402.8$, $p < .001$, $\eta p^2 = .88$ ($N = 54$, one participant lacking usable windows), identically with one and two decisions (window × task level $p = .87$). Across the full sample the effect between decisions persisted but was reduced by about 40% in millimeters, $F(1.60, 86.6) = 48.3$, $p < .001$, $\eta p^2 = .47$. The phase effect is of a different nature, $F(1.74, 93.9) = 112.1$, $p < .001$, $\eta p^2 = .67$: the number of alerts was identical across phases, the luminance of the outside scene was not. Pupil size also covaried negatively with near-display gaze within participants, $r = -.25$, $p < .001$.

Relations Among Indicators

Across participants, the fifteen correlations among effect scores are exploratory, none survives a Holm correction and they are not interpreted here. Within participants, once both the manipulation and each participant's own level are removed, perceived workload still covaried negatively with memory span, $r = -.29$, $p < .001$.

Indicators therefore differ not only in how strongly they respond but in what they respond to. Perceived workload tracks both factors. Near-display gaze tracks the task level upward and the flight phase downward, so on the phase factor the two dissociate. Memory span and the declared weight of piloting follow the flight phase far more than the task level ($\eta p^2 = .26$ vs. .03, and .29 vs. .10); the declared weight of the decisions shows the opposite profile (.04 vs. .18); decision performance responds to neither.

DISCUSSION

Within the same participants and the same manipulation, the indicators displayed the three relationships described by Hancock (2017): perceived workload, gaze and declared allocation associated with demand; decision performance remained insensitive to both factors, and memory performance to the task level except where capacity was still available; pupil diameter reversed. Read together they do not converge on a single estimate of load: they answer different questions.

Sensitivity of Perceived Workload

Perceived workload was the only indicator to rise with both factors, which invites reading it as the most valid of the six. But the manipulation was also the most visible thing in the experiment. Adding a first decision raised the

composite by 8.4 points; adding a second raised it by only 2.5, although the number of decisions doubled at that step. What weighed was the presence of a second task more than its volume. What the ratings shared with behavior was narrow: within participants, once the manipulation and each participant's own level were removed, they covaried with recall, r = –.29, and with neither near-display gaze nor pupil diameter. The composite reports experienced demand; it does not measure the resources a task consumes.

Limits of Stable Decision Performance

The stability of decision performance is the least informative element of this pattern. Performance was stable at a modest level: a fifth of decisions went unanswered, and accuracy on those answered stayed between .62 and .70. Little room was left for degradation, so a floor of difficulty remains a plausible reading, and the case for deliberate protection rests not on performance but on two allocation indicators: SWORD ratings show effort moving toward the decisions when they doubled, withdrawn from piloting or, under autopilot, from memorization (Yeh & Wickens, 1988), and near-display gaze rose with task level outside the decision windows themselves, a shift that was observed and not only declared.

Recall as an Index of Spare Capacity

Among the performance and physiological indicators, recall was the only one that responded to the demand of flying: gaze tracked the displays and fell when the aircraft demanded attention, decision performance moved with neither factor, and pupil diameter followed the light. Span fell by nearly two letters from autopilot to air defense, on a scale that needs no calibration, letters out of eleven above a known chance level. Its response to the secondary task, carried by the number of letters produced, was confined to autopilot, the condition in which the method is supposed to work: the flight task left capacity to spare, and the added decisions consumed it. A concurrent task indexes the capacity that remains only while the operator still serves it, and the level at which it stops responding to added load marks the point where that capacity is exhausted.

Two reservations apply. The phase effect is inseparable from the progression through the session, and the nine lists were presented in a fixed order, so proactive interference accumulates exactly as flight demand increases; a purely cumulative account, however, cannot produce the rise in recall across the three air-defense blocks. And the design cannot say which channel the cost fell on, since every alert carried a tone and, at the two-decision level, fell on the letter grid. A memorization task on a channel the other tasks leave free, and a flight demand kept below saturation, would let the index be read directly, which is what recommends it for further study.

Gaze and Pupil in a Luminous Cockpit

Near-display gaze had the largest task-level effect of the study, and its sensitivity has a clear source: the decisions were displayed in the cockpit, so the measure partly reflects the visual time-sharing the secondary task imposes.

That about a quarter of the effect survives when no map is displayed, more than a third in autopilot, indicates that it is not reducible to reading time. Across flight phases it moved the other way, falling as the aircraft demanded attention outside, which indicates what the measure captures: a direct index of how visual resources are shared between the outside scene and the cockpit (Wickens, 2008), not a covert marker of an internal state. An allocation index and a demand index need not agree, and in this study they did not.

The most plausible account of the pupillary result is optical, not cognitive: the decision task was itself a source of light, each decision beginning with a brief alert and continuing with a bright map that had to be inspected. Three observations are consistent with this reading: the constriction was concentrated within decision windows, was identical with one and two decisions, and covaried with the time spent looking at the displays. Two of them, however, do not discriminate between accounts: the cognitive demand of a single window is also identical whether one or two decisions occur, and near-display gaze entails convergence at less than a meter, which constricts the pupil through the near reflex.

The methodological consequence holds under either account. What reaches the retina depends on where the operator looks, and that depends on the task; current headsets do not record the luminance received at the eye. Until gaze-contingent photometry is accessible, pupil diameter should be used in immersive settings only for contrasts in which the visual event is physically identical.

A Question-Based Framework

These results suggest selecting measures by the question each can answer, and by the factor it answers it for (Table 2). Task performance is interpretable only alongside indicators of effort and the level at which it is stable. The task an operator gives up first indicates the capacity that remains, while it is still served. A physiological index first requires checking that the environment lets it reflect cognitive demand at all. What this adds to a general call for multimodality is an order, which question each measure answers, for which factor, and which cannot be interpreted in this environment.

Table 2: What each measure answered, and in which direction, for each of the two factors.

Question	Measure	With Task Level	With Flight Phase
Is demand rising?	Perceived workload	rises	rises
Is the operator coping?	Decision performance	no change, modest level	no change
How is effort allocated?	SWORD weights	toward the decisions	toward flying
	Near-display gaze	rises	falls
What capacity is left?	Recall	falls in autopilot only	falls
Is the index interpretable?	Pupil diameter	constricts (optical)	follows scene luminance

Limitations

The fixed order of phases and task levels is the principal limitation: both factors are confounded with time on task, with practice and, for the memorization task, with proactive interference across nine successive letter lists. The fatigue item shows this drift to be substantial, and the trends in decision accuracy and the improvement in ceiling conformity are equally compatible with practice or, for the latter, a change in strategy. Near-display gaze is a bounded proportion with a floor in the two demanding phases, where an analysis of variance is only an approximation, and it indexes any fixation inside the cockpit.

CONCLUSION

In immersive fighter-flight simulation, workload indicators associated, remained insensitive, or even reversed within the same participants and the same manipulation. Perceived workload rose with both factors; gaze tracked the task level and moved the opposite way across flight phases; decision performance held at a modest level; recall registered the demand of flying and, where capacity was still available, the added decisions; pupil diameter most plausibly reflected the light emitted by the decision task. Crossing flight phase with secondary-task level within participants is what made these divergences visible: read against a single factor, several of these indicators would have appeared to agree. None of them, alone or combined, yields a single estimate of workload, and the study does not claim one. What it offers instead is a map of what each indicator answers, for which factor, and under which conditions it can be read at all (Table 2), together with a candidate index of the capacity that remains: a concurrent memorization task, interpretable within the range where the operator still serves it. Three refinements would sharpen this reading: decoupling task order from flight phase, a decision task with room to degrade, and photometry at the eye. The design already separates the demand of flying from the demand of working the displays, the division a two-seat crew makes and a single-seat pilot carries alone, so reading each indicator against each factor opens a way to compare workload by role.

ACKNOWLEDGMENT

Many thanks to all the participants for their time and cooperation.

REFERENCES

Cegarra, J., & Chevalier, A. (2008). The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. Behavior Research Methods, 40(4), 988–1000.

Dhengre, S., & Rothrock, L. (2025). Investigating mental workload across task modalities: A multimodal analysis using pupillometry. Ergonomics, 68(9), 1458–1471.

Hancock, P. A. (2017). Whither Workload? Mapping a Path for Its Future Development. In L. Longo & M. C. Leva (Eds.), Human mental workload: Models and applications (Communications in Computer and Information Science, Vol. 726, pp. 3–17). Springer.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload (Advances in Psychology, Vol. 52, pp. 139–183). North-Holland.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. Science, 154(3756), 1583–1585.

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. Behavior Research Methods, 51(3), 1336–1342.

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. Frontiers in psychology, 13, 883321.

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. Behavior Research Methods, 55(6), 3055–3077.

Mohanavelu, K., Poonguzhali, S., Adalarasu, K., Ravi, D., Vijayakumar Chinnadurai, Vinutha, S., Ramachandran, K., & Srinivasan Jayaraman. (2020). Dynamic cognitive workload assessment for fighter pilots in simulated fighter aircraft environment using EEG. Biomedical Signal Processing and Control, 61, 102018.

Vidulich, M. A., Ward, G. F., & Schueren, J. (1991). Using the subjective workload dominance (SWORD) technique for projective workload assessment. Human Factors, 33(6), 677–691.

Wickens, C. D. (2008). Multiple resources and mental workload. Human Factors, 50(3), 449–455.

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. Human Factors: The Journal of the Human Factors and Ergonomics Society, 30(1), 111–120.