

Physiological and Subjective Assessment of Fighter Pilot Mental Workload

Anna J. C. Reuten; Juliette Bruin; Wietse D. Ledegang; Eric L. Groen; Mark M. J. Houben;
Erik-Jan Cornelisse; Charelle Bottenheft

- INTRODUCTION:** Next-generation air force operations increasingly depend on advanced flight systems, which raises questions about their impact on aircrew workload. Physiological measures enable continuous, objective, and unobtrusive assessments, overcoming some limitations of self-report measures. We investigated whether electroencephalogram, electrocardiogram, and electrodermal activity predicted subjective Rating Scale of Mental Effort (RSME) ratings of fighter pilots during an operationally relevant mission in an F-16 simulator.
- METHODS:** The mission included three flight segments expected to induce increasingly higher workload, with a mathematical task of two arithmetic difficulty levels added to each segment. This approach allowed us to explore whether physiology distinguished workload both across and within flight segments.
- RESULTS:** Subjective ratings confirmed an increase in workload throughout the mission (Δ RSME = 22). However, our manipulation of arithmetic difficulty was less effective: neither subjective ratings nor arithmetic performance significantly differentiated simple and complex calculations (Δ RSME = 5; Δ reaction time = 0.02 s, Δ unsuccessful trials = 0.5). None of our physiological measures—frontal theta power (Δ 0.08, Δ -0.07 mV), parietal alpha power (Δ -0.16, Δ 0.14 mV), heart rate (Δ 0.01, Δ -0.02 bpm), skin conductance response (Δ 0.0, Δ 0.0 mS), and skin conductance peaks (Δ 0.3, Δ 0.2)—systematically distinguished flight segments or arithmetic difficulty, respectively. Physiology also did not significantly predict subjective workload ($R^2 = 2.2\%$).
- DISCUSSION:** Large interindividual variability, use of workload management strategies, and relatively small task demand differences may explain the null results. Our study underscores the complexity of workload assessment in an operational setting and offers methodological insights that may inform future ecologically valid research.
- KEYWORDS:** psychophysiology, aviation, mental effort, workload, aircrew.

Reuten AJC, Bruin J, Ledegang WD, Groen EL, Houben MMJ, Cornelisse E-J, Bottenheft C. *Physiological and subjective assessment of fighter pilot mental workload*. *Aerosp Med Hum Perform*. 2026; 97(8):592–600.

The next generation air force is characterized by incremental usage of advanced communication, navigation, surveillance, and defense flight systems.^{1,2} These (automated) systems can benefit aircrew performance by providing critical information about threats, engagements, and friendly forces. However, they also introduce potential risks that may threaten flight safety, operational efficiency, and performance. For example, the abundance of available support systems could overload aircrew with information or system handling.^{1,3} Also, when these systems take over substantial aspects of the workload, vigilance might deteriorate, which could lead to missing of critical information and reduced situational awareness.⁴ These challenges are closely related to the concept of “mental workload.”

Mental workload has been the topic of ergonomic research for several decades, with applications not only in the field of aviation, but also in many other branches of transport ergonomics.⁵ There is no single agreed upon definition of it. We interpret mental workload as the subjective experience of task

From Human Performance, TNO, Soesterberg, The Netherlands; and Cranfield University, Cranfield, United Kingdom.

This manuscript was received for review in January 2026. It was accepted for publication in May 2026.

Address correspondence to: Anna J. C. Reuten, Ph.D., Kampweg 55, Soesterberg, Utrecht 3769 DE, Netherlands; anna.reuten@tno.nl.

Copyright © by The Authors.

This article is published Open Access under the CC-BY-NC license.

DOI: <https://doi.org/10.3357/AMHP.6875.2026>

demands, influenced by factors such as time pressure, contextual factors, and experience.^{5,6} Previous research has proposed an inverted U-shaped relationship between mental workload and task performance.^{5,7} According to this model, optimal performance is characterized by a situation in which there is neither overload nor underload; in other words, when an optimum is achieved between task demands and the operator's resources.

Mental effort is a key component of mental workload, representing the amount of invested cognitive resources during a task. If a person chooses not to invest effort in a task with high demands, the experienced workload may be low, despite its objective load. Effort thus plays a crucial role in maintaining performance, as individuals can increase their effort to compensate for increasing task demands.^{8–10} In such cases, performance may remain stable, masking the underlying cognitive strain. Therefore, performance alone is not a sufficient indicator of workload. To gain a more accurate understanding of mental workload and the invested effort, subjective assessments are collected via self-report rating scales such as the Rating Scale of Mental Effort (RSME)¹¹ or NASA Task Load Index.¹² However, such assessments are limited to discrete time points, prone to response biases, and may interfere with task execution. Therefore, it may be useful to estimate mental workload continuously, objectively, and unobtrusively.² Physiological responses can provide insight in brain activity and activation or suppression of the sympathetic and parasympathetic nervous system, all expected to vary with mental workload.^{13,14} Prior studies have successfully used psychophysiology to estimate levels of mental workload in the laboratory.^{6,15} The main aim of our study is to investigate the potential of physiology to predict subjective mental workload during a semicontrolled simulated flight mission.

Previous studies on physiological measurements of mental workload in pilots often used a limited set of physiological measures, primarily focusing on cardiovascular responses.^{3,6,16} These responses are relatively easy to obtain and provide minimal interference for the subject. However, they may not be sensitive enough to distinguish between flight tasks expected to evoke different levels of mental workload.¹⁷ Here we will extend cardiovascular responses with measurements of skin conductance and brain activity. Furthermore, prior studies did not always include subjective assessments to benchmark the interpretation of physiological responses.^{18–20} This can be problematic as flight tasks with higher workload may be confounded with increased muscle activity, which affects physiology.²¹ We therefore complement physiological recordings with subjective workload ratings to examine whether both indicators show comparable patterns. Moreover, we will investigate whether our measures can distinguish between different levels of workload across flight segments, as well as within individual segments. We will do so by adding a secondary mathematical task to each segment, consisting of audio recordings of simple and complex arithmetic calculations. Using this methodology allows us to investigate whether possible differences in physiology between segments reflect true variations in mental workload rather than confounding factors such as task-related artifacts.

Our study focuses on answering three research questions. We first investigate whether the flight segments and arithmetic difficulty levels within each segment induced differing levels of mental workload as indicated by subjective ratings and arithmetic performance. Secondly, we investigate whether the segments and arithmetic difficulty levels elicited different physiological responses. Lastly, we investigate whether our physiological measures can predict subjective mental workload.

METHODS

Subjects performed a flight mission in which they were tasked with detecting, intercepting, and eliminating a target within hostile territory while they performed secondary easy or complex arithmetic calculations. Throughout the mission, we recorded electroencephalogram (EEG), electrocardiogram (ECG), and electrodermal activity (EDA). At regular intervals, we asked subjects to provide a rating of mental workload using the RSME. We quantified arithmetic performance based on reaction time and the number of unsuccessful trials.

Subjects

Due to the specialized sample of F-16 pilots, we were restricted in the number and gender of subjects that we could include. Our sample consisted of 14 male pilots with an average of 2059 (SD = 699) flight hours in an F-16. All pilots except for three had flown an aircraft in the prior 10 d. Subjects were currently employed as a military fighter pilot or instructor ($N = 4$), air force brigadier general ($N = 1$), civil pilot or instructor ($N = 6$), other ($N = 1$), or were retired ($N = 2$). Their average age was 52 yr (SD = 9). We asked subjects to refrain from drinking alcohol or using drugs 12 h prior to the experiment. Approval for this study was granted by TNO's Institutional Review Board (reference number: 2024-072). All subjects provided written informed consent.

Equipment and Materials

We conducted the experiment using a fixed-base simulator of a Lockheed Martin F-16 Flying Falcon with an F-16 aerodynamic model (**Fig. 1**; multiSIM BV, Soesterberg, Netherlands). The interior of the simulator consisted of an original F-16 cockpit equipped with standard information displays, instrument panels, and HOTAS controls. The high-fidelity F-16 flight simulation model was created in D-SIM (multiSIM BV) and projected via a $240^\circ \times 140^\circ$ vertical projection dome. Head-up display symbology was projected as an overlay in the out-the-window visual. Subjects used the trim button on the control stick to respond to the mathematical task (described below).

The mission scenario was designed by a veteran F-16 weapon school instructor and experimental test pilot. It followed a semiscripted structure, starting with a “holding” segment (waiting in allied territory before crossing the Forward Line of Own Troops), followed by an “intercept” segment (crossing the Forward Line of Own Troops and monitoring a hostile aircraft via radar), and ending after an “engagement”



Fig. 1. The experimental set-up of the fixed-based F-16 cockpit, its projection dome, with a subject (left) and experiment leader sitting at the control station (right).

segment (eliminating the hostile aircraft and locating a ground target). The veteran pilot anticipated mental workload to increase throughout the mission, with the lowest levels expected during the holding segment and highest during the engagement segment.

To account for possible confounding factors associated with higher workload (e.g., muscle activity), we added a secondary auditory mathematical task with two difficulty levels to each flight segment, allowing us to compare workload under similar task conditions. The task consisted of simple vs. (more) complex arithmetic calculations alternated in four blocks of 30 s. Each block included six calculations with a maximum allowed response duration of 2 s. The task of subjects was to indicate whether the value of a two-operator calculation was higher or lower than 5 using a button press. The difficulty level of the calculations was manipulated by: 1) the number of digits that needed to be processed; and 2) the sign of the operators. Simple calculations always started with 0, meaning that subjects only needed to process the last two digits (e.g., $0 + 7 - 1$). Additionally, the first operator in the simple calculations was always a plus sign. Complex calculations consisted of three nonzero digits, with each calculation containing one plus and one minus operator in pseudorandom order, such that subjects could not predict the result before hearing the full equation. For both difficulty levels, restrictions ensured that the result of the calculations ranged from 1 up to 9 and was never equal to 5. No calculations were repeated (see **Appendix A** in the online version of this article for further details). We expressed performance on the mathematical task in two outcome variables: average reaction time in seconds and the number of unsuccessful calculations. Unsuccessful calculations are defined as the sum of incorrect and missing trials. In our calculation of reaction time, we imputed the maximum response time (2 s) for missing trials.

After each 30-s block within the mathematical task, subjects rated their perceived mental effort using the RSME.¹¹ The RSME scores range from 0–150, with higher scores indicating greater perceived workload or mental effort. The scale includes nine descriptors along the axis, such as “not effortful” at value 2

and “rather effortful” at value 58. We interpret this measure as a proxy for subjective mental workload.

Physiological assessments of workload were based on the data recorded during execution of the mathematical task. We recorded EEG, ECG, and EDA using the ActiveTwo AD-box (BioSemi BV, Amsterdam, Netherlands). For EEG, we used a 32-channel wet electrode cap. For ECG, we used one electrode positioned on the right clavicle and another on the lowest floating left rib. For EDA, we placed the electrodes on the middle and index finger of the right hand as it was least dominant for task execution. For each flight segment, we pooled the data from the two 30-s blocks of each mathematical difficulty level, yielding 60-s epochs from which all parameters described below were extracted.

For EEG, we first bandpass filtered the raw data with a Hamming windowed Sinc FIR filter (3–100 Hz) and removed powerline noise (50 Hz). To detect bad channels, we used the Python PyPREP package.²² We then fitted an Artifact Subspace Reconstruction algorithm to a (5-min) baseline recording taken at the beginning of the experiment. Artifact Subspace Reconstruction is a method that removes artifacts from EEG by comparing the data to a clean baseline. It detects unusual activity and replaces it with a cleaned version based on clean (i.e., baseline) brain patterns. Lastly, we performed an independent component analysis using MNE-Python's implementation²³ to identify signal sources. The resulting components were automatically classified with the ICLabel algorithm, which assigns labels such as “brain”, “eye blink”, and “muscle” based on their spatial, spectral, and temporal features.²⁴ Because of the noisy nature of the experiment (e.g., lesser control of task manipulation, more muscle activity compared to a laboratory setting, and a large amount of equipment in close proximity), we selected only the data with the label “brain” for further analysis. Based on visual inspection, all data was of reasonable to good quality, which meant that no excessively noisy channels were identified after preprocessing. To analyze power bands, we transformed the data for each channel into power spectral density estimates with use of MNE-Python.²³ Our frequency bands of interest were theta (4–7 Hz) and alpha (8–12 Hz) in the frontal (AF3, F3, Fz, F4, AF4) and parietal (P3, Pz, P4) channels, respectively. We restricted our analyses to these two frequency bands to reduce the number of statistical comparisons, given prior evidence of their particular sensitivity to indices of mental workload.²⁵ We used an average of alpha and theta power across the respective channels as our outcome variables.

For ECG, we retrieved heart rate from the signal using the peak detection algorithm by Pan and Tompkins.²⁶ This algorithm applies a third order Butterworth bandpass filter to smooth the signal. We discarded data points above 200 and below 30 bpm and those deviating by more than 2 SDs from the mean. Three subjects with excessively noisy signals were omitted from analysis. We used the average heart rate as our outcome variable.

For EDA, we applied a third order Savitsky-Golay filter to the raw signal to filter out noise. To retrieve phasic and

tonic components, we used the Ledapy 1.2.1 package (Python Software Foundation, Beaverton, OR). For one subject (different from the three subjects excluded in the ECG analysis), excessively noisy data was identified, whom we therefore excluded from further analysis. We used the mean amplitude of the phasic component, the skin conductance response, and the number of significant skin conductance response peaks (amplitude threshold of $0.01 \mu\text{S}$) as outcome variables.

While visualizing the results for the physiological measures, we observed a data pattern that was in the opposite direction from our expectation: heart rate decreased rather than increased throughout the course of the experiment (**Fig. A1**, found in the online version of this article). We assume that this effect might reflect increased familiarity with the laboratory setting or simulator. Such time-on-task effects could camouflage changes in physiology that are related to differences in mental workload, especially when these differences are expected to be small. To account for this, we normalized our physiological measures relative to a measurement of 15 s taken prior to the start of each new flight segment (i.e., the end of the prior segment). We could not apply this procedure to the EDA measures due to its low signal frequency and the short baseline period (e.g., no significant skin conductance peaks). Moreover, separating the tonic component from the phasic signal effectively normalizes the phasic component already.

Procedure

The veteran pilot, here acting as an instructor pilot, briefed the subjects about the flight objective and tasks. They then received instructions on rating their mental workload using the RSME, of which a printed version remained available throughout the mission on their knee pad. After the briefing, we equipped subjects with all physiological sensors. Subjects were then seated inside the simulator and given 10–15 min to familiarize themselves with the simulator controls. They shortly practiced all elements of the mission with assistance from the instructor pilot. At the end of the familiarization period, subjects practiced performing the simple and complex mathematical task (two blocks of 30 s). We explained that they could not neglect the mathematical task, even if workload increased incrementally from the flight task. They then started the mission, which had a total duration of approximately 10–15 min. After each 30-s block of the mathematical task, subjects verbally rated their mental workload using the RSME without pausing the scenario (**Fig. 2**). The instructor pilot acted as the observer during the scenario and provided a debrief afterwards.

Statistical Analysis

To answer our first research question, we focused on the RSME and on the reaction time and number of unsuccessful calculations on the mathematical task. For each of these three outcome variables, we first investigated the effect of flight segment (holding, intercept, and engagement) using a multilevel linear regression analysis. Herein we can include subject and arithmetic difficulty of the mathematical task as random effect variables to account for individual differences and task-related variability.

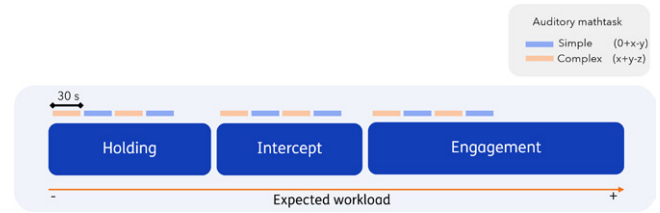


Fig. 2. Temporal overview of the flight segments and mathematical task. Subjects verbally provided a Rating Scale of Mental Effort (RSME) rating of subjectively experienced workload after each block of the auditory mathematical task. We analyzed the EEG, ECG, and EDA recordings taken during completion of the mathematical task. The order of flight segments was fixed; the order of the simple and complex arithmetic calculations was counterbalanced across subjects.

This approach allows for more accurate estimation of fixed effects, better generalization of results beyond our study sample, and is robust to assumption violations required by repeated measures ANOVA, such as sphericity. We compared the difference between the holding and intercept, the holding and engagement, and intercept and engagement segments. Secondly, we investigated whether arithmetic difficulty (simple vs. complex) influenced the three outcome variables within each flight segment. Therefore, we repeated the tests but included both segment and arithmetic difficulty as predictor variables. Here we included only subject as random effect variable. Out of all resulting pairwise comparisons, we were most interested in a comparison of the two arithmetic difficulty levels within each flight segment. We selected the holding segment as the reference for flight segment in both analyses, and the simple arithmetic calculations as reference for arithmetic difficulty in the second analysis. We used post hoc analysis to investigate a possible difference between the intercept and engagement segment in our first analysis.

To answer our second research question, we focused on five physiological outcome variables: frontal theta power (θ_{frontal}), parietal alpha power (α_{parietal}), heart rate, skin conductance response, and significant skin conductance peaks. For each of these variables, we performed the same analyses as for answering our first research question.

To answer our last research question, we performed a multilevel linear regression model with the RSME scores as outcome variable and all physiological measures as predictor variables. Here we implemented subject, segment, and arithmetic difficulty as random effect variables. For all analyses, we denoted significant effects as <0.05 .

RESULTS

First, we investigated whether the flight segments and arithmetic difficulty levels within each segment induced different levels of mental workload as indicated by subjective ratings and arithmetic performance. The pattern of results in **Fig. 3A** suggests that our manipulation of workload across segments was successful: on average, subjects ($N = 14$) provided increasingly higher RSME ratings throughout the flight mission. Workload increased from ratings reflecting “some effort” to

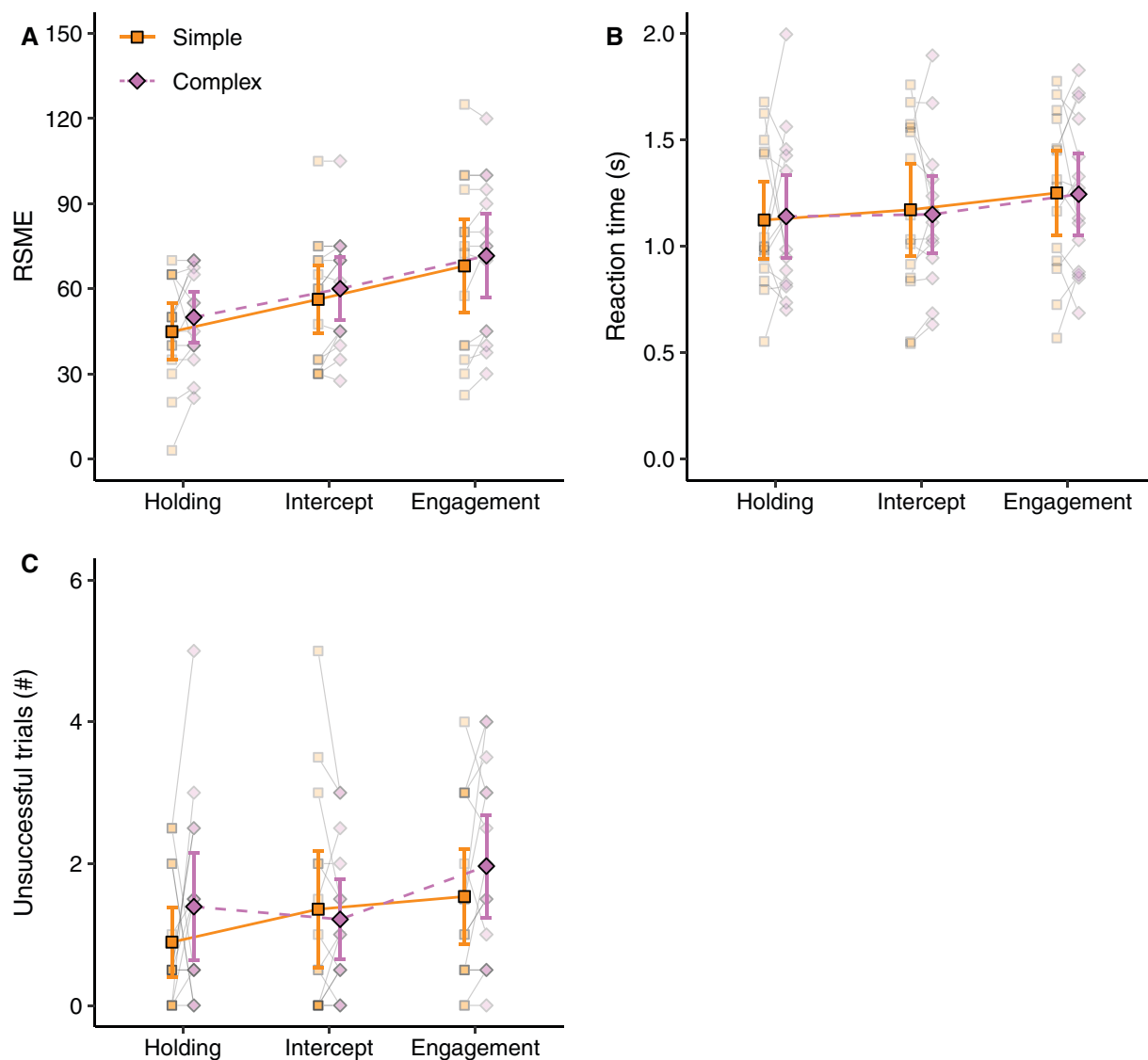


Fig. 3. Subjective and behavioral measures of mental workload. A) The self-reported workload based on the Rating Scale of Mental Effort (RSME). B) The reaction time for responding to one arithmetic calculation. C) The number of unsuccessful, either incorrect or missed (reaction time > 2 s), arithmetic calculations. In each panel, the solid data points represent an average across subjects and semitransparent data points the individual responses. Orange is for simple arithmetic calculations and pink for complex calculations. Error bars represent 95% confidence intervals.

about “great effort”. The results of the regression with flight segment as a predictor confirmed that the intercept [$B = 10.7$, $t(54) = 3.2$, $P = 0.002$] and engagement [$B = 22.4$, $t(54) = 6.7$, $P < 0.001$] segments evoked higher RSME ratings on average compared to the holding segment (Table AI, found in the online version of this article). The engagement segment also induced higher average RSME ratings than the intercept segment [$B = 11.7$, $t(54) = 3.5$, $P = 0.001$]. Our manipulation of the difficulty of the mathematical task was less effective. Fig. 3A shows that most subjects did provide higher RSME ratings for the complex relative to simple calculations within each flight segment. However, these differences were small and did not reach statistical significance in our regression with flight segment and arithmetic difficulty as predictors (Table AII, found in the online version of this article).

On average ($N = 14$), reaction time on the mathematical task slightly increased throughout the flight mission (Fig. 3B). The regression with flight segment as a predictor confirmed that the average reaction time significantly increased from the holding to the engagement flight segment [$B = 0.12$, $t(54) = 2.1$, $P = 0.041$]. The other two flight segment comparisons did not yield a significant difference (Table AIII, found in the online version of this article). The results of the regression with both flight segment and arithmetic difficulty as predictors suggest that the reaction time difference between the holding and engagement segments was driven by differences during the simple arithmetic task [Fig. 3B; $B = 0.5$, $t(65) = 2.3$, $P = 0.026$]. There was also a marginally significant difference between the holding and intercept segments for the simple calculations [$B = 0.4$, $t(65) = 2.0$, $P = 0.049$]. The pairwise

comparisons between arithmetic levels within each flight segment were not statistically significant (**Table AIV**, found in the online version of this article).

Fig. 3C ($N = 14$) demonstrates a small increase in the average number of unsuccessful calculations for the engagement compared to the holding and intercept flight segments. The results of the regression with flight segment as predictor partly confirmed this observation, with a significant difference between the engagement and holding segments [$B = 0.6$, $t(54) = 2.4$, $P = 0.022$]. Differences between the other flight segment comparisons did not reach significance (**Table AV**, found in the online version of this article). No systematic effect can be observed for the arithmetic difficulty levels within each flight segment (**Fig. 3C**). Indeed, the results of the regression with flight segment and arithmetic difficulty as predictors yielded no significant differences (**Table AVI**, found in the online version of this article).

Second, we investigated whether the flight segments and arithmetic difficulty levels elicited different physiological responses. The pattern of results in **Fig. 4A** ($N = 14$) suggests that average frontal theta power slightly increased from its baseline during the intercept segment, whereas it did not during the holding and engagement segments. The regression with flight segment as a predictor indicated a significant difference between the intercept and holding segments [$B = 0.2$, $t(54) = 2.4$, $P = 0.018$]. There were no significant differences in average theta power between the other two flight segment comparisons (**Table AVII**, found in the online version of this article). Further inspection of **Fig. 4A** suggests that theta power was comparable, though systematically higher on average, for the simple vs. complex arithmetic calculations within each flight segment. The regression with flight segment and arithmetic difficulty as predictors did not indicate any significant differences (**Table AVIII**, found in the online version of this article).

Fig. 4B ($N = 14$) suggests a marginal increase in the average parietal alpha power relative to the respective baselines of each flight segment throughout the flight mission. The results of the regression with flight segment as predictor indicated that alpha power was, on average, significantly lower in the engagement relative to holding flight segment [$B = -0.2$, $t(54) = -2.2$, $P = 0.032$]. Other pairwise comparisons were not significant (**Table AIX**, found in the online version of this article). The differences between simple and complex arithmetic calculations within each flight segment were small (**Fig. 4B**). None of the pairwise comparisons in the regression with flight segment and arithmetic difficulty as predictors yielded a significant difference (**Table AX**, found in the online version of this article).

Consistent with theta power, we also observed an increase in heart rate relative to its baseline for the intercept segment, but not for the holding and engagement segments (**Fig. 4C**; $N = 11$). The regression with flight segment as a predictor confirmed that average heart rate was significantly higher during the intercept compared to the holding segment [$B = 0.03$, $t(42) = 2.9$, $P = 0.005$] and showed a trend for an increase relative to the engagement segment [$B = 0.02$, $t(42) = 2.0$, $P = 0.057$]. There was no significant difference between the holding and

engagement segments (**Table AXI**, found in the online version of this article). Within each flight segment, we can observe that the average heart rate was systematically higher for the simple compared to complex calculations (**Fig. 4C**). However, the regression with flight segment and arithmetic difficulty as predictors indicated that these differences were small, not reaching statistical significance (**Table AXII**, found in the online version of this article).

The pattern of results in **Fig. 4D** ($N = 13$) does not suggest a systematic effect for the skin conductance response across flight segments or between the difficulty levels of the mathematical task within each segment. Neither regression analysis indicated significant effects (**Table AXIII** and **Table AXIV**, found in the online version of this article).

The number of average skin conductance peaks throughout the mission was about constant (**Fig. 4E**; $N = 13$). The regression with flight segment as a predictor indicated that none of the differences between flight segments reached significance (**Table AXV**, found in the online version of this article). Within each flight segment, the average number of skin conductance responses was comparable between the simple and complex arithmetic calculations (**Fig. 4E**). Accordingly, the regression analysis with segment and arithmetic difficulty as predictors yielded no significant effects (**Table AXVI**, found in the online version of this article).

Lastly, we investigated whether physiological responses predicted the subjectively rated mental workload. The results from the regression ($N = 10$) indicate that none of our physiological measures significantly contributed to predicting the RSME (**Table AXVII**, found in the online version of this article). In fact, collectively they explained only 2.2% of the variance in the provided RSME ratings.

DISCUSSION

Continuous, objective, and unobtrusive assessment of mental workload is relevant to individuals with safety-critical occupations such as aircrew. We here investigated whether EEG, ECG, and EDA can predict subjective mental workload during simulated flight. The mission scenario included three flight segments, with an additional mathematical task of two arithmetic difficulty levels added to each segment. Subjective ratings confirmed an a priori expected increase in mental workload throughout the flight mission. However, our manipulation of the difficulty level of the mathematical task did not significantly affect subjective ratings and arithmetic performance measures. None of our physiological measures—frontal theta power, parietal alpha power, heart rate, skin conductance response, and skin conductance response peaks—systematically distinguished flight segments or arithmetic difficulty. They also did not significantly predict the subjectively rated mental workload.

Considering that subjective workload and arithmetic performance did not significantly differ between the easy and complex arithmetic calculations, it is not that surprising that our physiological measures also failed to make this distinction.

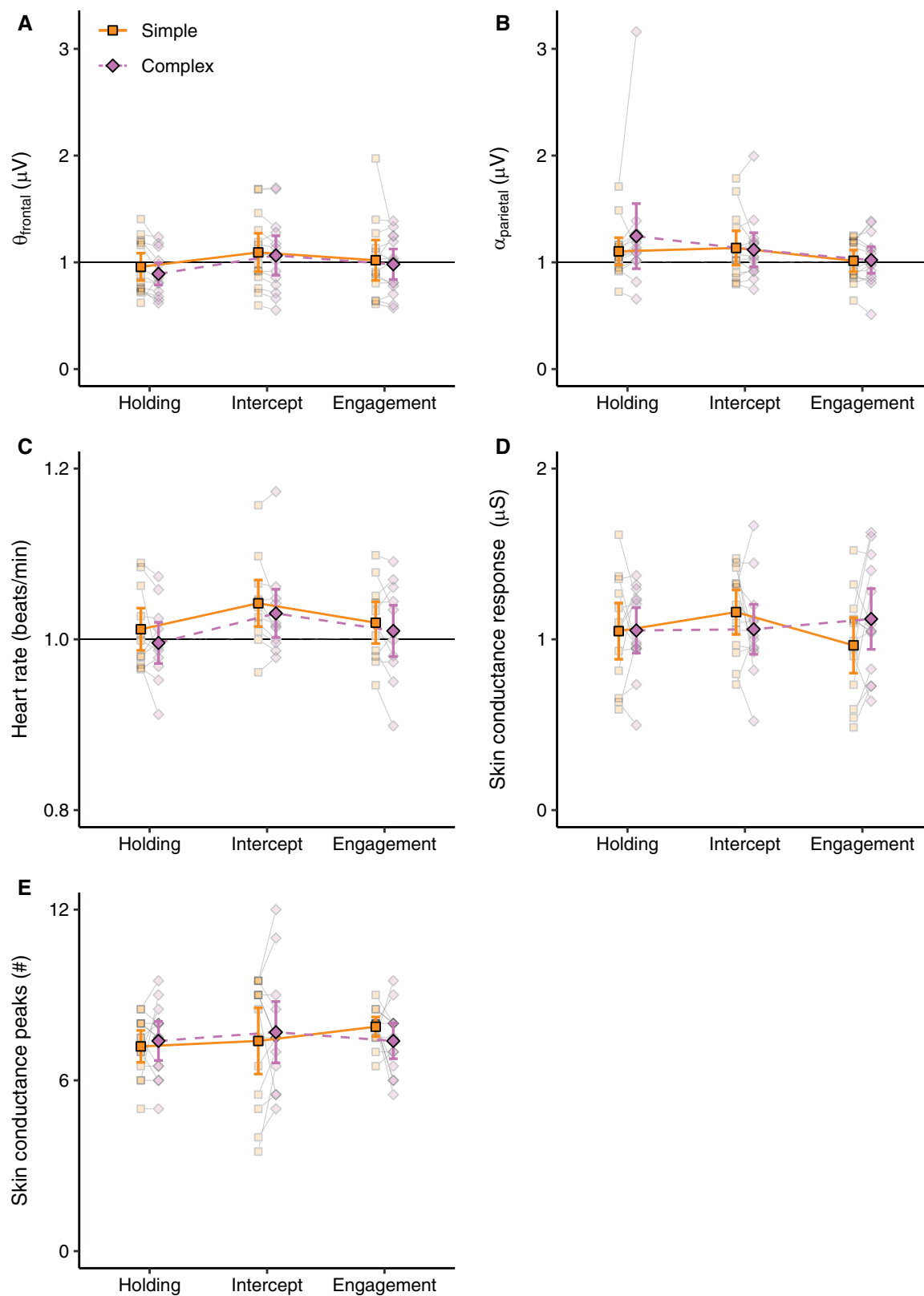


Fig. 4. In each panel, the solid data points represent an average of A) frontal theta power, B) parietal alpha power, C) heart rate, D) skin conductance response, and E) skin conductance peaks across subjects. The data presented in panels A-C are normalized relative to a baseline period taken directly prior to the start of each flight segment. See Fig. 3 for further details.

However, in contrast to previous studies,^{6,27,28} our physiological measures also did not vary with subjective ratings of workload across flight segments. This raises the question: how should we interpret this lack of an effect?

First of all, studies typically succeeded in distinguishing flight segments only in cases when task demands varied substantially. For example, cardiovascular measures differentiated takeoff and landing from midflight, but not takeoff from landing, considered only slightly more cognitively demanding.^{27,28} Similarly, Veltman¹⁷ reported that physiological measures could differentiate passive baseline from active flying, but could not differentiate between flight segments of expected different difficulty such as straight flight, complex turns, or responder tasks. Hogervorst *et al.*¹⁵ reported that the performance accuracy of classifier algorithms trained on (a combination of) physiological measures decreased when distinguishing between working memory tasks with more similar difficulty levels (2-back vs. 1-back or 1-back vs. 0-back) compared to more distinct ones (2-back vs. 0-back), and when using shorter data segments (30 s vs. 120 s). Differentiating subtle differences in levels of mental workload using physiological measures is challenging, particularly when relying on short measurement periods in an operational context. This challenge was compounded by the small sample size of our study due to using a highly specialized population of F-16 pilots. This increased the chance of making a type II error (i.e., false negative). However, studies that reported significant physiological differences across flight segments that varied substantially in task demands used similarly small samples (≤ 12 subjects).^{18,27,28} Nevertheless, we did not observe systematic patterns in the physiological data that pointed at insufficient statistical power. Our findings suggest that physiological measures may be less sensitive to variations in workload than subjective ratings, which did differ significantly across flight segments.

Second, pilots are trained in mental workload management to maintain stable performance levels under periods of high task demands.²⁹ Strategies such as task prioritization, automation, and externalization of memory cues are used to mitigate the risk of cognitive overload in critical situations such as close-range aerial combat or emergencies.¹ Less prioritized tasks are omitted and delayed until during periods of low workload, along with other preparatory tasks as to minimize underload.¹ Pilots should keep their workload below their maximum capacity, leaving spare resources to deal with unexpected situations such as unforeseen threats or system failures. Ayala *et al.*³⁰ reported that a secondary auditory oddball task did not affect flight performance of low-hour pilots during simulated flight, even though subjective and behavioral measures did suggest an increase in workload. While this self-regulation of mental workload is desirable and crucial for operational effectiveness and preserving flight safety, it may have obscured physiological markers of mental workload in our study. Subjective ratings did significantly differ across flight segments, which might be interpreted to reflect invested mental effort rather than the level of experienced workload when assuming efficient use of workload management strategies. The absence of a ground truth

makes it difficult to interpret the absence of significant differences in physiological measures as it could reflect a true lack in workload differences or be the result of insufficient sensitivity and measurement noise.

Lastly, the aspects raised bring us to consider a fundamental aspect: what component of workload is it we measure when assessing physiology? For example, Brouwer *et al.*⁹ demonstrated that physiological measures did not reflect reduced mental effort associated with learning to master a task, while at the same time they could differentiate between levels of task difficulty. In other words, physiology was sensitive to the overall level of task demands (i.e., a main effect of n-back load) but could not capture decreases in invested effort of the n-back task due to practice (i.e., no interaction of n-back load across learning blocks). This highlights the complex relationship between physiology and workload. In our study, subjective ratings pointed to the engagement segment as the most demanding part of the flight mission. However, for frontal theta power and heart rate we observed a pattern of results that suggested highest workload for the intercept segment. This apparent contradiction suggests that physiological responses may be driven by factors beyond perceived task demands. Post hoc interviews indicated that the intercept segment was associated with greater uncertainty about upcoming threats and a reduced sense of control compared to the engagement. Additionally, there is the anticipation of high workload for the engagement segment. Such factors may reflect “affective” rather than cognitive aspects of workload.³¹ These considerations highlight the complexity of predicting mental workload in an operational context: it is a multidimensional construct that yields noisy measurements. At the same time, physiology reflects a wide range of mental and bodily states that are linked to (non-specific) differences in neural activity (i.e., mainly EEG) and arousal (i.e., mainly ECG and EDA).

Even though we could not confirm the value of physiological measures for predicting subjective workload, our study had several strengths. We used an experimental design which included representative end users in an operationally relevant setting. Additionally, we included a secondary task with the aim of separating cognitive from physical components of workload, as to help isolate task-related confounds (e.g., muscle activity) that have been overlooked in prior studies. Furthermore, combining multiple physiological measures provided a more comprehensive approach to assessing mental workload since they reflect different underlying processes (i.e., neural activity vs. arousal). However, large interindividual variability in perceived workload and physiology, usage of workload management strategies to minimize the risk of cognitive over- and underload, relatively small differences in task demands, and nonspecificity of physiological responses have likely all contributed to a null result. This implies that studies require more statistical power to detect significant effects, which asks for strong experimental manipulations, large sample sizes, and long measurement periods. At the same time, this presents a clear challenge for real-time monitoring of workload using physiology; for example, to adjust training difficulty in simulated flight or to tailor information delivery and automation level in real flight.

Such applications might become feasible with the development of personalized models that are trained on large datasets. In conclusion, our study underscores the complexity of workload assessment in an operational setting and offers methodological insights that may inform future ecologically valid research.

ACKNOWLEDGMENTS

The authors would like to thank all subjects for contributing to this study.

Financial Disclosure Statement: This work was funded by the Dutch Ministry of Defence as part of Research Programme V2306. The authors have no competing interests to declare.

Authors and Affiliations: Anna J. C. Reuten, Ph.D., M.Sc., Juliette Bruin, M.Sc., B.Sc., Wietse D. Ledegang, M.Sc., Eric L. Groen, Ph.D., M.Sc., Mark M. J. Houben, Ph.D., M.Sc., Erik-Jan Cornelisse, M.Sc., and Charelle Bottenheft, Ph.D., M.Sc., Human Performance, TNO, Soesterberg, Netherlands; and Eric L. Groen, Professor, Cranfield University, Cranfield, United Kingdom.

REFERENCES

- Burian B, Pruchnicki S, Rogers J, Christopher B, Williams K, et al. Single-pilot workload management. Moffett Field (CA): NASA Ames Research Center; 2013. Report No.: ADA602396.
- Masi G, Amprimo G, Ferraris C, Priano L. Stress and workload assessment in aviation—a narrative review. *Sensors* (Basel). 2023 Mar 28; 23(7):3556.
- Mansikka HP. Fighter pilot's performance and mental workload [dissertation]. Coventry (United Kingdom): Coventry University; 2016.
- Young MS, Stanton NA. Attention and automation: new perspectives on mental underload and performance. *Theor Issues Ergon Sci*. 2002; 3(2):178–94.
- Young MS, Brookhuis KA, Wickens CD, Hancock PA. State of science: mental workload in ergonomics. *Ergonomics*. 2015; 58(1):1–17.
- Charles RL, Nixon J. Measuring mental workload using physiological measures: a systematic review. *Appl Ergon*. 2019 Jan; 74:221–32.
- Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci Biobehav Rev*. 2014 Jul; 44:58–75.
- Bottenheft C, Groen EL, Van Erp JBF. Experimental validation of a performance curve using a Tetris game. *BMC Psychol*. 2025 Jul 9; 13(1):755.
- Brouwer A-M, Hogervorst MA, Holeywijn M, Van Erp JBF. Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *Int J Psychophysiol*. 2014 Aug; 93(2):242–52.
- Mansikka H, Virtanen K, Harris D. Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Trans Human-Mach Syst*. 2019 Feb; 49(1):1–9.
- Zijlstra F. Efficiency in work behaviour—a design approach for modern tools [dissertation]. Delft, The Netherlands: Technische Universiteit Delft; 1993.
- Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. *Proc Human Factors Ergon Soc Annu Meet*. 2006; 50(9):904–8.
- Tao D, Tan H, Wang H, Zhang X, Qu X, Zhang T. A systematic review of physiological measures of mental workload. *Int J Environ Res Public Health*. 2019; 16(15):2716.
- Chikhi S, Matton N, Blanchet S. EEG power spectral measures of cognitive workload: a meta-analysis. *Psychophysiology*. 2022 Jun; 59(6):e14009.
- Hogervorst MA, Brouwer A-M, Van Erp JBF. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front Neurosci*. 2014 Oct 14; 8:322.
- Mohanavelu K, Poonguzhali S, Ravi D, Singh PK, Mahajabin M, et al. Cognitive workload analysis of fighter aircraft pilots in flight simulator environment. *Defence Science Journal*. 2020; 70(2):131–9.
- Veltman JA. A comparative study of psychophysiological reactions during simulator and real flight. *Int J Aviat Psychol*. 2002; 12(1):33–48.
- Dussault C, Jouanin J-C, Guezennec C-Y. EEG and ECG changes during selected flight sequences. *Aviat Space Environ Med*. 2004 Oct; 75(10):889–97.
- Lahtinen T, Koskelo J, Laitinen T, Leino T. Heart rate and performance during combat missions in a flight simulator. *Aviat Space Environ Med*. 2007 Apr; 78(4):387–91.
- Mansikka H, Simola P, Virtanen K, Harris D, Oksama L. Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics*. 2016 Oct; 59(10):1344–52.
- Sweeney KT, Ward TE, McLoone SF. Artifact removal in physiological signals—practices and possibilities. *IEEE Trans Inf Technol Biomed*. 2012 May; 16(3):488–500.
- Bigdely-Shamlo N, Mullen T, Kothe C, Su K-M, Robbins KA. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Front Neuroinform*. 2015 Jun 18; 9:16.
- Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, et al. MNE software for processing MEG and EEG data. *Neuroimage*. 2014 Feb 1; 86:446–60.
- Pion-Tonachini L, Kreutz-Delgado K, Makeig S. ICLabel: an automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage*. 2019 Sep; 198:181–97.
- Gevins A, Smith ME. Neurophysiological measures of cognitive workload during human computer interaction. *Theor Issues Ergon Sci*. 2003; 4(1–2):113–31.
- Pan J, Tompkins WJ. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*. 1985 Mar; 32(3):230–6.
- Lee Y-H, Liu B-S. Inflight workload assessment: comparison of subjective and physiological measurements. *Aviat Space Environ Med*. 2003 Oct; 74(10):1078–84.
- Scannella S, Peysakhovich V, Ehrig F, Lepron E, Dehais F. Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Hum Factors*. 2018 Nov; 60(7):922–35.
- Lubold N, Finseth T, Gorry T, Inderberg L, Pellerino J, Dodd S. Flight crew task management: baseline assessment. Phoenix (AZ): Honeywell Aerospace Advanced Technology; 2024. Contract No.: 693KA8-22-T-00008.
- Ayala N, Kearns S, Irving E, Cao S, Niechwiej-Szwedo E. The effects of a dual task on gaze behavior examined during a simulated flight in low-time pilots. *Front Psychol*. 2024 Nov 19; 15:1439401.
- Neerincx MA. Modelling cognitive and affective load for the design of human-machine collaboration. In: Harris D, editor. *Engineering Psychology and Cognitive Ergonomics*. Berlin Heidelberg (Germany): Springer Berlin Heidelberg; 2007:568–74.