JURNAL ILMIAH TEKNIK INDUSTRI
ISSN: 1412-6869 (Print), ISSN: 2460-4038 (Online)
Journal homepage: http://journals.ums.ac.id/index.php/jiti/index
doi: 10.23917/jiti.v23i2.5296

A Preliminary Investigation on Vicarious Observation of Mental Workload

Ridwan Aji Budi Prasetyo[1a*]

Abstract. *The primary objective of this study is to conduct an initial investigation into the possibility of reliably predicting mental workload (MWL) just from task observation. In order to accomplish the goal of this study, a completely repeated-measures design was implemented. Twenty-one participants were instructed to examine sampling videos of individual MATB and combined subtasks at two different levels of demand. Afterwards, participants were instructed to subjectively assess the MWL of the task using the NASA-TLX scale. The findings suggest that participants can differentiate the mental workload (MWL) of the system monitoring task, but not the other subtasks or the tasks when they are combined. The findings also indicate that the ratings for the subtask varied significantly between different degrees of pressure. The results can be elucidated via the perspective of signal detection and heuristics theories. This paper also addresses the methodological constraints and potential practical application.*

Keywords: *mental workload; MATB; NASA-TLX; observation*

I. INTRODUCTION

The concept of mental workload (MWL) has been around for more than 50 years (Moray, 1979). The concept, which originated in the 1970s, has been implemented in diverse contexts to enhance the reliability of human operatives engaging with technologies. It could be argued that the significance of MWL has grown in recent years due to the fact that digitization and automation have converted laborious tasks into cognitive ones (Sharples, 2019) (Yassierli et al., 2016). Additionally, the MWL concept is highly intuitive and consistent with our daily lives. It appears to be widespread in numerous work environments. For instance, we may encounter situations in which we must manage an excessive quantity of information or data while performing routine work duties; or we might find that our focus on the road while operating a motor vehicle hinders our ability to participate in a conversation with a fellow passenger (Sheykhfard et al., 2023). In addition, the evaluation MWL has progressively embraced a multidimensional strategy that incorporates objective and subjective metrics. This trend signifies a growing recognition that MWL is an intricate and multifaceted concept that defies precise quantification through a single metric. As an illustration, researchers have integrated conventional subjective rating scales, including the NASA Task Load Index (TLX) (Hart & Staveland, 1988), with the more sophisticated measurement techniques employing psychophysiological sensors including fNIRS (Ayaz et al., 2012), EEG (Aghajani et al., 2017), ECG (Mansikka et al., 2016), eye-tracker (Appel et al., 2018), or facial thermography (Marinescu et al., 2018). These objective evaluations are facilitated through the application of non-intrusive and cost-effective sensors. By incorporating various metrics, it is possible to acquire a holistic understanding of every aspect of MWL and the way they impact performance.

Nevertheless, the conventional emphasis of MWL research has been on the human operator executing tasks, especially when employing subjective assessment methods. Put simply, the evaluation of MWL involves inquiries made by human operators either during or immediately after the task completion. For instance, the NASA-TLX (Hart & Staveland, 1988) is typically administered after the completion of the entire task under consideration. If the tasks are

[1] Department of Psychology, Brawijaya University, Jalan Veteran, Malang, 65145, Indonesia

[a] email: ridwan.prasetyo@ub.ac.id
◆ corresponding author

Submited: 24-07-2024 Revised: 04-12-2024
Accepted: 15-12-2024

performed sequentially, it may be administered in between. While the NASA-TLX may have arguably withstood the test of time (Hart, 2006), this approach has encountered criticism due to concerns that recall bias could result from post-task scale administration (Devos et al., 2020; Tingting et al., 2024). To address this limitation, MWL assessment could be practically administered while performing the task at hand, although NASA-TLX is unsuitable for the intended purpose. This scale comprises six dimensions (in the form of Likert-style queries) that require responses, potentially causing a significant disruption to the ongoing task. Instantaneous Self-Assessment of Workload instrument (ISA) was created as a potential substitute for addressing this challenge. The ISA is a subjective technique utilised to evaluate the MWL encountered during a task in a timely manner (Brennan, 1992). Originally, its purpose was to quantify the MWL of air traffic controllers (ATC). The instantaneous nature of the scale makes it less intrusive and more appropriate for evaluation in real-time. The perceived MWL of the operator is evaluated using a five-point rating scale, where a score of "1" denotes a low workload and a score of "5" signifies a high workload. Throughout a task, the scale is administered at various intervals, including every two minutes (Kirwan et al., 1997) or every 45 seconds (Marinescu et al., 2018).

The idea of estimating MWL prior to the actual task is therefore still rarely explored in previous studies. Few studies, however, sought to distinguish between retrospective (post-task) and prospective (pre-task) evaluations of MWL in the performance of medical surgery tasks (Sublette et al., 2009, 2010). Based on the results of their experiments, it appeared that participants' expectations regarding the difficulty of the task vary. It was stated that this was dependent upon the dimensions of MWL under consideration. In addition, prospective evaluation of MWL can function as a reliable approximation for retrospective evaluation in their situation, where physical and temporal demands constitute the main components of the overall workload. A similar study examined a comparable construct but obtained differing results (Sublette et al., 2009). It was discovered that users' perceptions of task difficulty varied depending on the type of task (e.g., increasing or decreasing) between before and after completing the task. This remained constant, specifically, in the dependent-structured task. It decreased during the parallel-structured or designated task, while it increased as the unnamed task progressed. Such research can yield valuable insights regarding participants' perceptions of the task and its various components prior to the actual task. As a result, it may be possible to identify more effective strategies or methods for accomplishing the actual task.

Predicting MWL prior to task execution is therefore important due to its relation to performance, either directly or indirectly. A study from (Nuutila et al., 2021) suggested that task performance was negatively affected by the initial perception of task difficulty and a larger rise in difficulty over time. Furthermore, a study from (Maynard & Hakel, 1997) explained that subjective task complexity partially mediated the effects of objective task complexity and cognitive ability on task performance. Thus, the task performance was diminished because of the perception that the task was complex rather than straightforward, which was a result of the high objective task complexity and low cognitive ability. It is possible that the negative correlation between task performance and perceptions of complexity is mediated by self-efficacy, resulting in a decrease in one's confidence in the successful completion of the task. In addition, from a learning perspective, it is argued that when someone engages in a learning task, they form metacognitive perspectives of the present learning situation. Their perceptions are influenced by their prior knowledge of similar activities, including past processes and their outcomes (Efklides, 2008).

Based on these arguments, the present study aims to preliminarily investigate whether MWL can be accurately predicted by merely observing the task. The novelty of this study lies in the endeavour to test this rarely explored concept. Furthermore, the results gained from this study may be exploited as evidence to introduce the

task prior to its actual execution for several contexts and purposes, e.g. training or skill acquisition.

II. RESEARCH METHOD

Theoretical Frameworks

To the author's best knowledge, there is no single theory to date specifically addressing this phenomenon. However, several overarching theories can be employed to offer approximate explanations, such as signal detection theory or heuristics and biases.

Signal detection theory (SDT) provides a theoretical framework used to analyse the behavioural reactions of humans when undertaking a perceptual task in a laboratory setting (Hautus, 2015). The theory posits that a subject's capacity to differentiate between sensory inputs is constrained by the fluctuation in the mental representations of those stimuli. The response to a stimulus might vary from one occurrence to another due to differences in the stimulus itself and/or random fluctuations in the neurological system. If there is an overlap in the distributions of these representations for two separate stimuli, then it is certain that some errors will occur. As the degree of overlap between the distribution increases, the number of mistakes produced also increases (Sumner & Sumner, 2020). Specifically in human factors studies, SDT offers a method for differentiating between accuracy and criterion setting in decision-making contexts. This is beneficial for assessing the efficacy of decision-making capabilities exhibited by an intelligent machine, a human user, or a human-machine system. SDT is valuable for determining the optimal allocation of subtasks and roles in human-computer monitoring systems (Parasuraman et al., 2000; Parasuraman & Wisdom, 1985). SDT can be implemented in the form of a multitasking environment that enables different stimuli presentation at the same time (Kim et al., 2016). The stimuli itself can be in the form of visuospatial (Gugerell et al., 2024), auditory (Moseley et al., 2016), or tactile (Boldt et al., 2014).

While SDT focuses on the stimuli presentation and the way humans can discriminate between stimuli, the understanding of human judgement is also essential. Human judgment or decision-making arguably involves varying degrees of irrationality (Pothos et al., 2021) because of the existence of cognitive biases and heuristics (Kahneman & Frederick, 2002). Although there are several versions of these theories, they all share the idea that human decision making involves two distinct processes: System 1 and System 2. System 1 is characterised by its fast and instinctive nature, which can make it susceptible to mistakes. On the other hand, System 2 is a more deliberate and controlled process that depends on logical reasoning. The utilisation of either one or both of these systems depends on the number of cognitive resources and time that each generally requires (Dehdashti et al., 2020). The use System 1 in prospective observation can lead to somewhat accurate results. A study from (Cabrera et al., 2015) in the context of predicting patient's sickness, for example, claimed that the decision-making process of System 1, which relies on limited information, demonstrated a sensitivity of around 80% in accurately predicting acuity (whether patient is sick or not) and disposition (whether patient should be dispositioned to home care or ICU). However, its performance was comparatively poorer in predicting ICU admission and diagnosis. The study suggested that System 1 decision-making is not enough for making final judgments in these areas, but it probably serves as a cognitive foundation for System 2 decision-making.

Experiment task and study design

This study was conducted using a fully repeated-measures design, that is, all participants in this study participated in all conditions. In other words, the design was using two-factor with repeated measures on both factors. The independent variable was the demand levels and types of multi attributes subtasks as generated by the Multi-Attribute Task Battery (MATB), a platform to test operator's multitasking performance designed and developed by NASA

Table 1. The details of the stimuli presentation

Level	TRACK	SYSMON	RESMAN
Low demand	Low preset default	Two deflections per minute	One pump fails every minute
High demand	Medium preset default	30 deflections per minute	One or two pumps fail for 15 seconds every minute

Table 2. The characteristics of participants on this study based on age range, gender, and employment status

Characteristics		Frequency	Percentage
Age range	25-30	8	38%
	31-35	7	33%
	36-40	6	29%
Gender	Male	12	57%
	Female	9	43%
Status	Student	16	76%
	Staff	5	24%

(Comstock & Arnegard, 1992). The demand levels comprised low and high, which was based on predetermined stimuli presence during the task execution. In general, low demand task consisted of less presentation of stimuli, while high demand task comprised more rapid stimuli presentation. The types of tasks in this experiment were classified into four categories, namely, system monitoring (SYSMON), resource management (RESMAN), tracking (TRACK), and combined (MULTI). The SYSMON task involved monitoring changes within a system and responding to these changes accordingly. The RESMAN task simulated a fuel management task. It required participants to maintain the designated level of fuels from two tanks and identify errors in pumping system that may occur during the trial. The TRACK task involved maintaining the position of a target within a specific boundary. The MULTI task included all MATB tasks together in one frame. We followed the validation work from Kennedy and Parker (Kennedy & Parker, 2017) in determining the frequency and the timing of the stimuli. Table 1 shows the details of the stimuli presentation for both low and high demand tasks.

The dependent variable of this study was the perceived evaluation of workload as measured by NASA-TLX (Hart & Staveland, 1988). The raw score of the scale was chosen as the primary approach for measuring MWL because of its simplicity. Furthermore, studies have found that the sensitivity of the original NASA-TLX scale may be higher, the same, or lower than that of the scale's weighted version (Hart, 2006). The reason for using weighted or unweighted techniques in NASA-TLX scoring is unclear, indicating that users can choose the method of scoring that is most suitable for their study.

Participants

A total of twenty-one participants were selected from the pool of university students and staff (Mage = 32.43, SDage = 4.50) using several recruitment methods, including institutional emails and communication groups that contained the study link. In addition, a personalised strategy was implemented by personally reaching out to colleagues and peers to ensure a maximum number of participants. The study received ethics approval from the Faculty of Engineering Research Ethics Committee at the University of Nottingham. Table 2 specifies the characteristics of participants.

Materials

MATB task videos. The tasks from Multi-Attribute Task Battery version 2 (MATB-II) as previously described were recorded in a single video for each task and demand levels, resulting in eight separate videos (SYSMON, RESMAN,

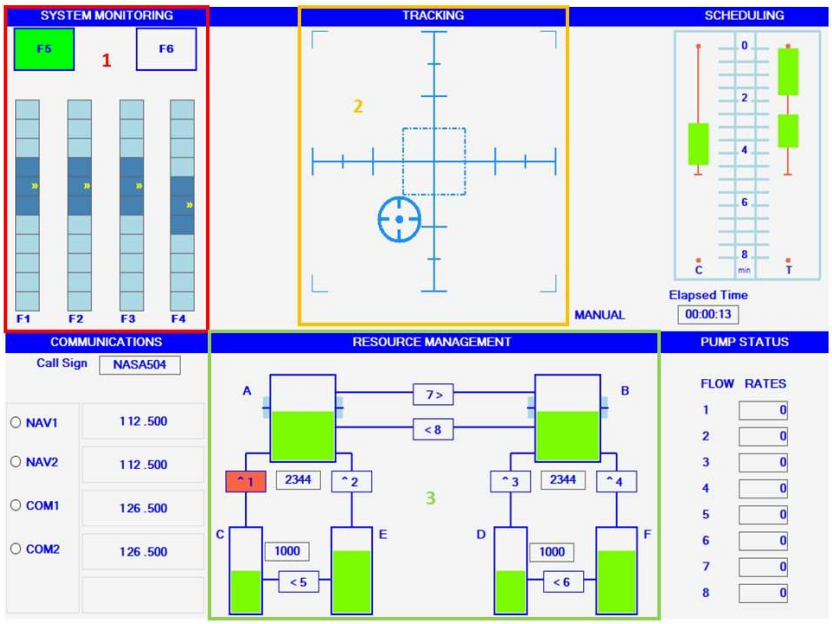

Figure 1. The default interface of MATB (combined tasks); The red rectangle (1) is system monitoring task, the amber rectangle (2) is tracking task, and the green rectangle (3) is resource management task.

TRACK, and MULTI tasks; low and high demand for each task). Each video lasted for one minute. The videos were recorded on an MP4 format with 1280 x 720 30fps resolution. Since MATB interface showed all the tasks in one single frame, each individual task was manually and proportionally cropped so that participants were able to focus on the task in question. The videos were stored and presented through Microsoft Forms. Figure 1 shows the interface of MATB and its subtasks.

NASA-TLX scale. The NASA-TLX scale was recreated using Microsoft Forms from its standard paper-and-pencil form. The scale was administered immediately after the appearance of the videos. A total of eight NASA-TLX were administered. Participants were required to respond to the scale after the conclusion of the video on each task.

Procedure

This was an online experiment using a sharable link. Once the study had been advertised, participants were able to access a URL and immediately directed to a Microsoft Forms page. Participants were required to read an information sheet and sign an informed consent. After indicating their agreement to participate in this study, several demographic questions were presented, followed by an instruction page. The experiment session commenced with the presentation of low-demand MATB task videos that followed a specific sequence, starting from SYSMON, TRACK, RESMAN, and MULTI tasks. The sequence of low-demand tasks was succeeded by a sequence of high-demand MATB tasks, which were identical in order and protocol. Participants were instructed to watch the video thoroughly and asked to complete NASA-TLX questions before proceeding to the next video. Monetary compensation was then provided as a reward for their participation in the study. Figure 2 shows the sequence of this experiment.

Hypothesis and statistical analysis approach

This experiment aimed to test if participants were able to distinguish between high and low demand MATB tasks merely based on their observation of the sampled videos of the tasks. Therefore, the (null) hypothesis statement for this experiment was:

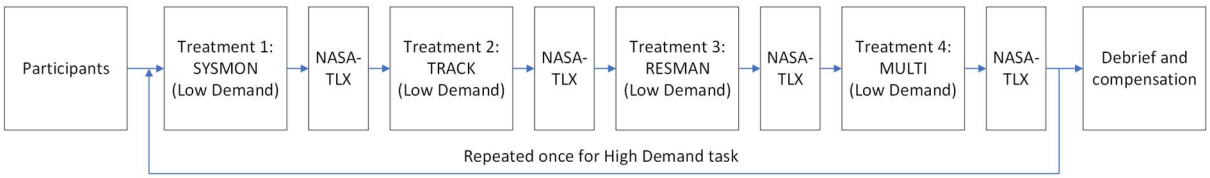

Figure 2. The sequence of the experiment; SYSMON = system monitoring, TRACK = tracking, RESMAN = resource management, MULTI = combined tasks

H1: There will be no differences in the total raw score of NASA-TLX after observing a video of high-demand task, compared to after observing a video of low-demand task, in all MATB subtasks (SYSMON, TRACK, RESMAN, and MULTI).

H2: There will be no differences in the total raw score of NASA-TLX for all MATB subtasks within certain demand level (low or high).

To test the hypothesis, we applied a 2 (low vs. high demand) x 4 (SYSMON, TRACK, RESMAN, and MULTI) repeated measures ANOVA. The factorial design for the ANOVA examined the demand levels and the total score of NASA-TLX in each MATB subtasks. Regarding the ANOVA results, our primary focus was on identifying any significant interaction at first, i.e. between demand levels and MATB subtasks. If an interaction effect was detected, pairwise comparisons were conducted using Bonferroni correction to identify the differences. The statistical analysis was performed mainly using R software (R Core Team, 2021)

The fully repeated-measure ANOVA model for analysing the data is expressed in the equation of:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \rho_i + \gamma_{ij} + \eta_{ik} + \varepsilon_{ijk} \quad \text{..... (1)}$$

Y_{ijk} is the subjective MWL score (NASA-TLX total score) taken at certain demand levels (k) from a MATB subtask (j) in subject number i. α_j is the main effect of demand levels subject to $\Sigma\alpha_j = 0$. β_k is the main effect of MATB subtask subject to $\Sigma\beta_k = 0$. $(\alpha\beta)_{jk}$ is the interaction effect of demand levels and MATB subtasks. ρ_i is the main effect of subjects $\sim N(0, \sigma_i^2)$. γ_{ij} is the interaction effect of subjects and demand levels $\sim N(0, \sigma_{ij}^2)$. η_{ik} is the interaction effect of subjects and MATB subtasks $\sim N(0, \sigma_{ik}^2)$. ε_{ijk} is the error term $\sim N(0, \sigma^2)$.

III. Result and Discussion

Interaction effect between demand levels and MATB subtasks

Prior to the analysis, several assumptions underlying ANOVA were checked. The results from Shapiro-Wilks normality test suggested that the data from TRK task was not theoretically obtained from a normal distribution. The data was then transformed using the Box Cox method. The process involved determining the optimal lambda value by a linear regression model and subsequently applying the following formula to convert the data.

$$y' = \frac{y^{\lambda}-1}{\lambda} \quad \text{..... (2)}$$

After the data was transformed, a two-way ANOVA was conducted to test the interaction effect. The results suggested that there was no statistically significant two-way interactions between demand levels and MATB subtasks, $F(2.31, 46.17) = 1.527$, $p = 0.260$. Since we did not find a significant interaction, the subsequent ANOVA procedure was to interpret the main effects for the demand levels and MATB subtasks. As seen in Table 3, the main effect of demand levels ($F(1, 20) = 4.679$, $p = 0.030$) and MATB subtasks ($F(3, 60) = 50.348$, $p = 0.000$) were statistically significant.

Main effect of demand levels

A post-hoc test was conducted to specifically find the interaction by testing the simple main effect of demand levels on subjective MWL score as measured by the total score of the NASA-TLX.

Table 3. ANOVA table (type III tests)

Effect	DFn	DFd	F	*p*
Task	3	60	50.348	0.000*
Level	1	20	4.679	0.030*
Task:Level	2.31	46.17	1.527	0.260

Table 4. Descriptive statistics and pairwise test for the NASA-TLX total score.

Task	Demand levels	Mean	SD	*p* value
SYSMON	Low	4.881	2.134	0.035*
	High	5.817	1.733	
TRACK	Low	3.722	1.654	0.108
	High	4.167	1.978	
RESMAN	Low	4.944	1.967	0.303
	High	5.230	2.172	
MULTI	Low	7.246	1.558	0.235
	High	7.579	1.459	

*$p < 0.05$

The effect of demand levels was statistically significant in system monitoring task (SYSMON) (p = 0.035), but not in tracking (TRACK) (p = 0.108), resource management (RESMAN) (p = 0.303), and combined tasks (MULTI) (p = 0.235). Table 4 shows the mean differences in the NASA-TLX total score. The results of the test suggested that participants can distinguish between low and high MWL after observing the sampled video of the system monitoring task. For the rest of the MATB subtasks, including the combination of these subtasks, the results implied that these tasks were not distinguishable by just observing the sampled task videos. Based on these results, hypotheses H1 can be partially rejected, suggesting the difference was partially found.

Main effect of MATB subtasks

A post-hoc test was conducted to specifically find the interaction by testing the simple main effect of MATB subtasks on subjective MWL score as measured by the total score of the NASA-TLX. The effect of MATB subtasks were all statistically significant except in SYSMON and TRACK tasks during both low (p = 1.00) and high (p = 0.972) demand levels. These results suggested that, after observing the sampled videos of MATB subtasks in either demand level, participants scored MWL in each subtask differently. In both demand levels, TRACK task was scored as the lowest compared to SYSMON and RESMAN, which tended to be similar, and MULTI task as the highest. Based on these results, hypotheses H2 can be partially rejected, suggesting the difference was partially found. Figure 3 shows the differences of MWL score between MATB subtasks in both demand levels.

Discussions

From our study, the results suggest that the system monitoring task was the only MATB subtask whose demands were possible to be predicted by merely seeing the prospective task. The tendency may arise from the discrepancy in stimuli of the MATB tasks, which are a type of 'signal detection' test that necessitates participants to respond appropriately. According to a research conducted by (Everly, 2016), tasks involving signal detection are more easily distinguished when there is a significant difference between the stimuli. During high-demand system monitoring tasks, one can see fast fluctuations in the task, characterised by alternating green and red lights and four scales that rapidly ascend and descend. To achieve success in this task, it is necessary to promptly hit the appropriate keys on the keyboard during the real-life simulation of MATB. The inherent nature

of the task may lead individuals to perceive it as mentally and physically challenging, thus resulting in frustration during actual performance. The task feeling of "rushing" when reacting to the assignment was similarly difficult to comprehend until really experienced.

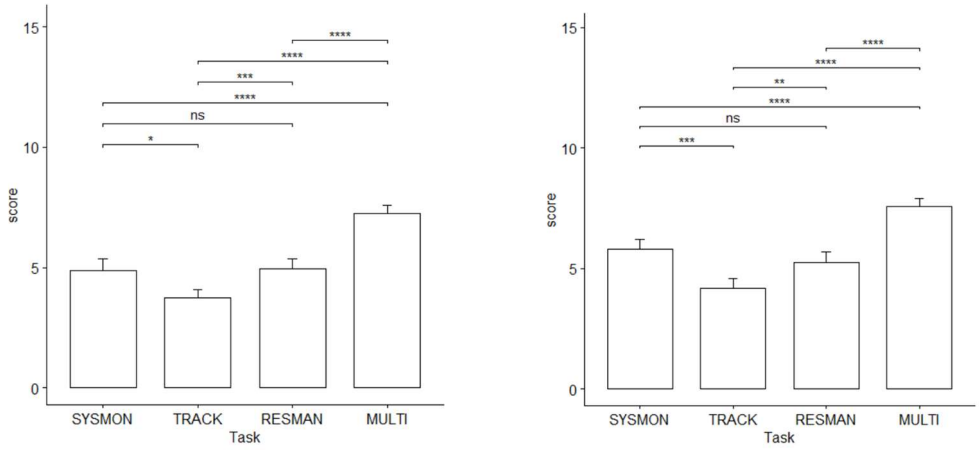

Figure 3. The differences of MWL score between MATB subtasks in both demand levels (left = low, right = high); * = significant at p < 0.05; ** = p < 0.01; *** = p < 0.01; **** = p < 0.0001

at hand was deemed to have a substantial MWL.

In contrast to a system monitoring tasks that involves multiple features, the tracking task just focuses on the movement of the target cursor when it departs its designated region. The intended outcome of this task was to adjust and maintain control of the joystick (usually a joystick) to reach the designated region. The nature of this task may not provide a significant sense of cognitive demand or frustration, as would be the case with a system monitoring task. Although using the joystick may require physical exertion, participants appeared to have difficulty envisioning the experience. This might be due to the joystick control being unfamiliar to most participants. Even in everyday situations, like gaming, a controlling device such as a joystick may not be essential. The lack of capacity to perceive the physical aspect of the activity may be correlated with effort and performance scores in all subtasks and the overall task. In the NASA-TLX brief explanation, the "effort" dimensions were defined as the level of difficulty involved in completing the task, while "performance" referred to the degree of success achieved in the task at hand. These two factors seem difficult to understand as they lack first-hand knowledge with the task. Just like the physical aspect, the

On the other hand, the task of resource management necessitates a more extensive cognitive process, as participants must compute the appropriate allocation of fuels across tanks, while also dealing with intermittent pump failures throughout the session. Thus, resource management appears to be the one subtask in MATB that does not yield an instant predicted response. Participants' responds would vary due to the task's necessity for employing multiple strategies, with participants having total freedom over how they applied these strategies. Observing the video on resource management may not effectively convey the sense of urgency in addressing variations in demand. In addition, the work of monitoring resources appears to be less dynamic when compared to the chores of monitoring and tracking the system. Participants were unable to differentiate between different degrees of task demand in terms of all elements of the NASA-TLX, including the overall score of MWL.

Regarding the combined tasks (MULTI), it may be stated that the task cannot be distinguished based on the levels of task demand. One possible reason for this outcome is that combining all subtasks has caused the differences between the parts of the tasks to become less

noticeable. Put simply, the capacity to differentiate between these specific tasks was causing confusion between them. The video of the MULTI task may have a range of stimuli that are slightly distinct from each other. These stimuli might help explain the lack of variation in outcomes across different degrees of task demand. Based on the current discussion, it can be inferred that even if the hypothesis is only partially supported, we cannot accurately forecast mental workload (MWL) in advance during a task. Without first-hand experience, it is impossible to accurately evaluate the level of difficulty or complexity of a task unless the aspects of the task are significantly different from each other.

While the abovementioned discussion focuses on the participant's prediction from the perspective of signal detection feature, understanding the way humans process partial information and generate predictions was also essential. From heuristics and biases theory, System 1 is often employed to generate predictions. More specifically, anchoring might be employed as a technique to support a judgment by accessing particular facts provided (Furnham & Boo, 2011). When attempting to estimate conditions that are unknown, individuals frequently rely on readily available information and progressively refine their estimate until a reasonable approximation is achieved. Anchoring refers to the phenomenon where an estimate is influenced by an initial value, causing it to be biased towards that value and resulting in premature changes (Lee & Hamilton, 2022). Related to this study, participants utilised specific information from the sampled videos of the MATB subtasks, and determined whether the task at hand was low or high MWL. For example, as discussed earlier, rapid changes of features in high demand system monitoring task might be used by participants as an anchor to conclude that the task was difficult, thus, having higher MWL. Nevertheless, from our study, MATB subtasks other than SYSMON failed to serve as an anchor for most participants. It can be implied that different anchors might be used by participants to predict the MWL of the subtasks. This conclusion was supported by the results from this experiment suggesting that there were significant differences in the prediction between the MATB subtasks during either demand levels.

This study was evidently preliminary, necessitating more investigation into several problems, such as the influence of expectations on strategy selection, the effects of experience and individual variations, and the alignment between prospective and retrospective judgments. We also recognise the constraints of our study, namely in relation to the methodology and the extent of its coverage. Future research might consider expanding the scope by doing this study with subject-matter experts. Conducting both prospective and retrospective evaluations of MWL in a controlled laboratory setting might be advantageous for assessing the accuracy and consistency of the expected results. Despite its exploratory nature and limitations, the study has yielded useful insights on the vicarious observation of MWL.

IV. CONCLUSION

This study finds that the viewing of sampling videos of a task can partially predict the MWL of the task. The MWL of the system monitoring task in the MATB may be precisely determined by analysing the sample video of the task. However, the MWL of the other subtasks in MATB and the combined subtasks cannot be reliably differentiated. Our study also discovered that the subjective MWL varied across different subtasks and demand levels. Specifically, the tracking task had the lowest MWL score, followed by the resource management and system monitoring tasks, which were very close. The resource management task had the highest MWL score. The findings of this study indicate that the availability of cues in each subtask might influence how participants anticipate the mental workload for the related task. Additionally, this study suggests that participants may utilise distinct information from the stimuli in each subtask to formulate their forecast of the MWL. Our study was exploratory and hence requires more exploration into several aspects. Although the study is exploratory and has limits, it has

provided valuable insights on the vicarious observation of MWL.

REFERENCES

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring Mental Workload with EEG+fNIRS. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00359

Appel, T., Scharinger, C., Gerjets, P., & Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1–8. https://doi.org/10.1145/3204493.3204531

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*(1), 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023

Boldt, R., Gogulski, J., Gúzman-Lopéz, J., Carlson, S., & Pertovaara, A. (2014). Two-point tactile discrimination ability is influenced by temporal features of stimulation. *Experimental Brain Research*, *232*(7), 2179–2185. https://doi.org/10.1007/s00221-014-3908-y

Brennan, S. (1992). *An experimental report on rating scale descriptior sets for the instantaneous self assessment (ISA) recorder* (DRA Technical Memorandum (CAD5) 92017). DRA Maritime Command and Control Division. https://skybrary.aero/bookshelf/books/1963.pdf

Cabrera, D., Thomas, J. F., Wiswell, J. L., Walston, J. M., Anderson, J. R., Hess, E. P., & Bellolio, M. F. (2015). Accuracy of 'My Gut Feeling:' Comparing System 1 to System 2 Decision-Making for Acuity Prediction, Disposition and Diagnosis in an Academic Emergency Department. *Western Journal of Emergency Medicine*, *16*(5), 653–657. https://doi.org/10.5811/westjem.2015.5.25301

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (NASA-TM-104174, NAS 1.15:104174). NASA Langley Research Center. https://matb.larc.nasa.gov/files/2014/03/Comstock-Arnegard-Original-TM.pdf

Dehdashti, S., Fell, L., & Bruza, P. (2020). On the Irrationality of Being in Two Minds. *Entropy*, *22*(2). https://doi.org/10.3390/e22020174

Devos, H., Gustafson, K., Ahmadnezhad, P., Liao, K., Mahnken, J. D., Brooks, W. M., & Burns, J. M. (2020). Psychometric Properties of NASA-TLX and Index of Cognitive Activity as Measures of Cognitive Workload in Older Adults. *Brain Sciences*, *10*(12), Article 12. https://doi.org/10.3390/brainsci10120994

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277–287. https://doi.org/10.1027/1016-9040.13.4.277

Everly, J. J. (2016). Human Performance on a Signal Detection Task: Discriminability and Sensitivity to Reinforcement. *The Psychological Record*, *66*(1), 139–151. https://doi.org/10.1007/s40732-015-0159-7

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

Gugerell, D., Gollan, B., Stolte, M., & Ansorge, U. (2024). Studying the Role of Visuospatial Attention in the Multi-Attribute Task Battery II. *Applied Sciences*, *14*(8), Article 8. https://doi.org/10.3390/app14083158

Hart, S. G. (2006). Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

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* (pp. 139–183). North-Holland Press. http://archive.org/details/nasa_techdoc_20000004342

Hautus, M. (2015). Signal Detection Theory. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 946–951). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.43090-4

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In D. Griffin, D. Kahneman, & T. Gilovich (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Kennedy, L., & Parker, S. H. (2017). Making MATB-II medical: Pilot testing results to determine a novel lab-based, stress-inducing task. *Proceedings of the 2017 International Symposium on Human Factors and Ergonomics in Health Care*, *6*, 201–208. https://doi.org/10.1177/2327857917061044

Kim, J. H., Yang, X., & Putri, M. (2016). Multitasking Performance and Workload during a Continuous

Monitoring Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 665–669. https://doi.org/10.1177/1541931213601153

Kirwan, B., Evans, A., Donohoe, L., Kilner, A., Lamoureux, T., Atkinson, T., & MacKendrick, H. (1997, June 16). Human Factors in the ATM System Design Life Cycle. *Human Factors in the ATM System Design Life Cycle Report*. FAA/Eurocontrol ATM R&D Seminar, Paris.

Lee, J., & Hamilton, J. T. (2022). Anchoring in the past, tweeting from the present: Cognitive bias in journalists' word choices. *PLoS ONE*, *17*(3). https://doi.org/10.1371/journal.pone.0263730

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016). Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics*, *59*(10), 1344–1352. https://doi.org/10.1080/00140139.2015.1136699

Marinescu, A. C., Sharples, S., Ritchie, A. C., Sánchez López, T., McDowell, M., & Morvan, H. P. (2018). Physiological Parameter Response to Variation of Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *60*(1), 31–56. https://doi.org/10.1177/0018720817733101

Maynard, D. C., & Hakel, M. D. (1997). Effects of Objective and Subjective Task Complexity on Performance. *Human Performance*, *10*(4), 303–330. https://doi.org/10.1207/s15327043hup1004_1

Moray, N. (Ed.). (1979). *Mental Workload: Its Theory and Measurement*. Springer US. http://link.springer.com/10.1007/978-1-4757-0884-4

Moseley, P., Smailes, D., Ellison, A., & Fernyhough, C. (2016). The effect of auditory verbal imagery on signal detection in hallucination-prone individuals. *Cognition*, *146*, 206–216. https://doi.org/10.1016/j.cognition.2015.09.015

Nuutila, K., Tapola, A., Tuominen, H., Molnár, G., & Niemivirta, M. (2021). Mutual relationships between the levels of and changes in interest, self-efficacy, and perceived difficulty during task engagement. *Learning and Individual Differences*, *92*, 102090. https://doi.org/10.1016/j.lindif.2021.102090

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy Signal Detection Theory: Basic Postulates and Formulas for Analyzing Human and Machine Performance. *Human Factors*, *42*(4), 636–659. https://doi.org/10.1518/001872000779697980

Parasuraman, R., & Wisdom, G. (1985). The Use of Signal Detection Theory in Research on Human-Computer Interaction. *Proceedings of the Human Factors Society Annual Meeting*, *29*(1), 33–37. https://doi.org/10.1177/154193128502900113

Pothos, E. M., Waddup, O. J., Kouassi, P., & Yearsley, J. M. (2021). What Is Rational and Irrational in Human Decision Making. *Quantum Reports*, *3*(1), https://doi.org/10.3390/quantum3010014

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software].

Sharples, S. (2019). Workload II: A Future Paradigm for Analysis and Measurement. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (pp. 489–498). Springer International Publishing. https://doi.org/10.1007/978-3-319-96071-5_52

Sheykhfard, A., Haghighi, F., & Das, S. (2023). How does talking with passengers threatens pedestrian life? An analysis of drivers' performance based on real-world driving data. *Transportation Research Part F: Traffic Psychology and Behaviour*, *95*, 464–479. https://doi.org/10.1016/j.trf.2023.05.010

Sublette, M., Carswell, C. M., Grant, R., Klein, M., Seales, W. B., & Clarke, D. (2009). Anticipated vs. Experienced Workload: How Accurately Can People Predict Task Demand? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/154193120905301846

Sublette, M., Carswell, C. M., Grant, R., Seidelman, W., Clark, D., & Seales, B. (2010). Anticipating Workload: Which Facets of Task Difficulty are Easiest to Predict? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(19), 1704–1708. https://doi.org/10.1177/154193121005401972

Sumner, C. J., & Sumner, S. (2020). Signal detection: Applying analysis methods from psychology to animal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1802), 20190480. https://doi.org/10.1098/rstb.2019.0480

Tingting, P., Xun, D., Jun, W., Shu, D., & Shan, Z. (2024). Mediator Effects of Cognitive Load on Association between Self-Efficacy and Task Load in Intensive Care Unit Nurses. *Journal of Nursing Management*, *2024*, 1–7. https://doi.org/10.1155/2024/5562751

Yassierli, Y., Aisha, A. N., & Nugraha, A. G. (2016). Pengembangan Alat Pengukuran Kelelahan Mental Berbasis Uji Flicker. *Jurnal Teknik Industri: Jurnal Keilmuan Dan Aplikasi Teknik Industri*, *18*(1), 11–20. https://doi.org/10.9744/jti.18.1.11-20