2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 6-10, 2024. Sarawak, Malaysia

Effects of Sensor Setup Time and Comfort on User Experience in Physiological Computing*

Vesna D. Novak, *Senior Member, IEEE*, Robert A. Kaya, Collyn J. Erion, Mohammad Sohorab Hossain, *Student Member, IEEE,* and Joshua D. Clapp

Abstract— **Physiological sensors are commonly applied for user state monitoring and consequent machine behavior adaptation in applications such as rehabilitation and intelligent cars. While more accurate user state monitoring is known to lead to better user experience, increased accuracy often requires more sensors or more complex sensors. The increased setup time and discomfort involved in the use of such sensors may itself worsen user experience. To examine this effect, we conducted a study where 72 participants interacted with a computer-based multitasking scenario whose difficulty was periodically adapted – ostensibly based on data from either a remote eye tracker or a lab-grade "wet" electroencephalography sensor. Deception was used to ensure consistent difficulty adaptation accuracies, and user experience was measured with the Intrinsic Motivation Inventory, NASA Task Load Index, and an ad-hoc scale. We found few user experience differences between the eye tracker and electroencephalography sensor - while one interaction effect was noted, it was small, and there were no other differences. This result is at first surprising and seems to indicate that comfort and setup time are not major factors for laboratory-based user experience evaluations of such technologies. However, the result is likely due to a suboptimal study protocol where each participant interacted with only one sensor. In future work, we will use an alternate protocol to further explore the effects of user comfort and setup time on user experience.**

I. Introduction

Physiological computing is a term used to describe any technological system that measures human physiological data and either displays these data to the human or adapts its own functionality to them [1]. Such systems are useful in many biomedical fields. For example, in rehabilitation robotics, physiological responses such as heart rate are commonly used to estimate the patient's workload and adapt the exercise difficulty accordingly [2]. Similarly, in affect-aware learning, a technological system can use physiological responses to assess learner engagement and adapt the learning materials accordingly [3]. In intelligent cars, physiological measurements can be used to assess driver attention levels and alert the driver if they are distracted or engage driving assistance if they are overwhelmed [4]. As a more casual example, music listeners could use physiological responses to optimize song selections to better achieve desired mood states [5]. Other applications are described in, e.g., a review paper by Aranha et al. [6].

In physiological computing systems, the user's state is commonly inferred by applying classification, regression and other algorithms to diverse measurements: electrocardiography [7], photoplethysmography, skin conductance, respiration, electroencephalography (EEG) [8], functional near infrared spectroscopy [9] and others. These sensors differ in their complexity and application time: for example, skin conductance requires two reusable dry electrodes commonly placed on the fingertips and is easily self-applied in less than 30 seconds [10] while laboratory-grade EEG requires 5-30 minutes to apply, cannot be self-applied, and often requires either skin gel or a saline-soaked sensor cap, reducing participant comfort [11]. We may then ask: do the benefits provided by more complex equipment outweigh the additional setup time? For example, in our own prior study on biocooperative rehabilitation robotics, we evaluated the ability of a rehabilitation robot to classify the user's workload into two classes (low vs. high) with or without physiological measurements. We found that adding physiological measurements to the robot's standard sensors increased the workload classification accuracy from 81.8% to 89.4% but required a few minutes of setup time [2], which could have been spent simply performing rehabilitation exercises instead.

Historically, developers of physiological computing systems have primarily focused on improving the accuracy of user state recognition and consequent technology adaptation without considering how this accuracy interacts with other aspects of the system. Thus, researchers have specifically called for more studies into broader user experience with physiological computing technologies [12], [13]. Our team previously conducted a few studies where we systematically varied the technology's adaptation accuracy and measured subjective user experience with physiological computing systems [14], [15]. Those studies showed that user experience generally improves with adaptation accuracy and that users can perceive a difference between accuracies that differ by 10-15% [14], [15]. Thus, adding additional sensors seems justified as long as they increase the physiological computing system's adaptation accuracy by a reasonable amount. We may, however, ask a different question: if two physiological computing systems behave the same way but one relies on more complex sensors, will users prefer the simpler one? This would, for example, be a point in favor of consumer-grade EEG devices, which are significantly more comfortable

* This work was supported by the National Science Foundation, grant numbers 2007908 and 2151464.

V. D. Novak and M. S. Hossain are with the Department of Electrical and Computer Engineering, University of Cincinnati, Cincinnati, OH 45221 USA (phone: 513-556-4765; e-mails: novakdn@ucmail.uc.edu, hossaimo@mail.uc.edu).

R. A. Kaya, C. J. Erion and J. D. Clapp are with the Department of Psychology, University of Wyoming, Laramie, WY 82071 USA (e-mails: rkaya@uwyo.edu, cerion@uwyo.edu, jclapp@uwyo.edu).

2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC) | 978-1-6654-1020-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/SMC54092.2024.10831219

and have a faster setup time than lab-grade devices even if they do not achieve the same accuracy [16].

This paper presents a pilot study to compare user experience with two physiological computing systems: one based on a noncontact eye tracker and one based on a "wet" EEG sensor, with associated differences in setup time and discomfort. While different physiological computing systems commonly exhibit different behavior due to hardware and software differences, a deception-based protocol from previous work [15] was used to induce similar technology adaptation behavior in both systems. As part of the deception, all physiological data were ignored by the system, and adaptation was instead done by simply following the user's preference a predefined percentage of the time. Thus, the two systems did not differ in adaptation accuracy, and any differences in user experience were likely due to differences in comfort and setup time. Our research question was: Does a comfortable, easy-to-set-up physiological computing system result in a better user experience than a less user-friendly system?

II. Materials and Methods

A. Participants

Seventy-two participants were recruited among University of Wyoming students. There were 56 women, 15 men, and one agender participant. They were 19.5 ± 2.4 years old (mean ± standard deviation). Each participant was randomly assigned to one of four groups corresponding to the sensor (eye tracker or EEG) and the magnitude of difficulty adaptation actions (large or small – explained later). The "eye tracker & small" group had 21 participants while all other groups had 17 each. Students received either course credit or $15 for participation.

B. Sensors and Scenario

Each participant took part in a single session where they interacted with a computer-based multitasking scenario while "monitored" with either an eye tracker or EEG sensor depending on the participant's group. The eye tracker was the Gazepoint GP3 (Vancouver, Canada) remote eye tracker placed under the screen on which the scenario was presented. The EEG system was a Geodesic Sensor Net with 128 electrodes (Electrical Geodesics Inc., USA). Fig. 1 shows a person interacting with the scenario using both sensor types.

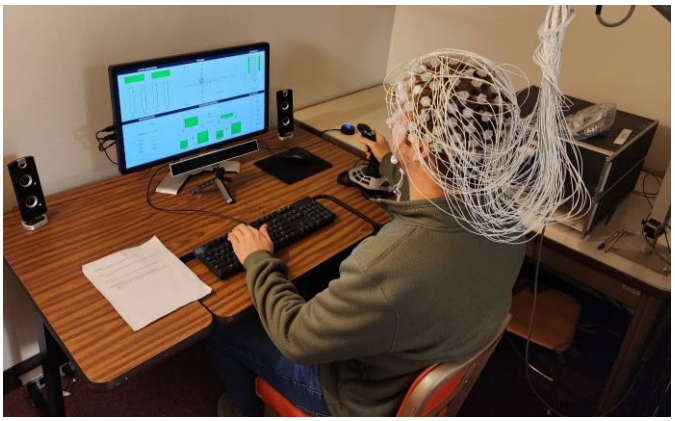

Figure 1. A person interacts with the adaptive OpenMATB scenario while monitored using both the electroencephalography sensor net and the eye tracker. Actual participants experienced only one sensor type.

The sensors were selected as two extremes in physiological computing – the eye tracker involves no physical contact and little setup time while the EEG involves potentially unpleasant physical contact and a much longer setup time.

The scenario (Fig. 2) was a modified version of the OpenMATB, a popular multitasking scenario used in physiological computing [17]. Participants interacted with it using a computer screen, speakers, keyboard, and joystick. The modified version was the same as in our previous study [15] and consisted of three subtasks:

- Tracking (Fig. 2, top center): A blue reticle starts at the center of the tracking section and gradually drifts toward the edges. The participant must use the joystick to keep it close to the center (indicated by a central 'target' square). If the reticle stays outside the center too long, it flashes red and an error counter on the screen increments by one.
- System monitoring (Fig. 2, top left): There are four vertical status indicator columns (F1-F4) and two warning lights (F5-F6). The columns have arrows in them that begin near the center, but occasionally move toward the top/bottom edge. If an arrow gets far away from the center, the participant must press the corresponding F1-F4 key on the keyboard to reset it. The warning lights periodically turn yellow, and the participant must then press the corresponding F5-F6 key. If the participant fails to press a required key or presses one of the F1-F6 keys at an unnecessary time, the corresponding indicator briefly turns red and the error counter increments by one.
- Communications (Fig. 2, bottom left): Periodically, a voice comes over the speaker that instructs a listener to change one setting (NAV1, NAV2, COM1, or COM2) to a specific number. The message includes an identifier, and the participant has their own identifier shown on the screen (e.g., DP94). If the identifier in the message does not match the participant's identifier, the participant does not need to take action. If the identifiers match, the participant should use the up/down keys to switch to the target setting and then use the left/right keys to decrement/increment the number to the requested value. If the participant fails to do so in time, the

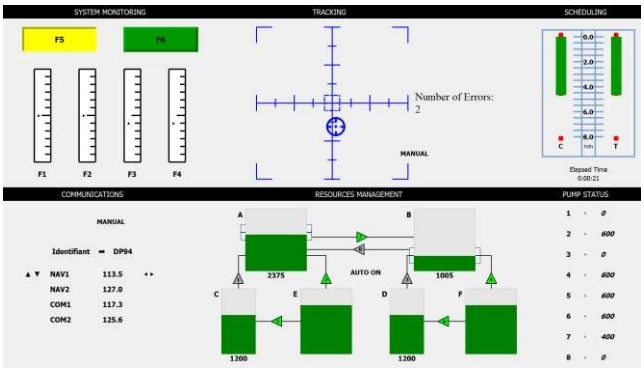

Figure 2. The adaptive OpenMATB used in this study. It was created for our previous study [15] based on an open-source implementation [17]. The user focuses on three tasks: system monitoring (top left), tracking (top middle), and communications (bottom left).

section flashes red and the error counter increments by one. If the participant changes the settings when not required, no penalty is given.

Additionally, three subtasks on the screen (scheduling, resources management, pump status) run automatically and do not need attention from the user. These can be controlled manually in the original OpenMATB [17] but were changed to run automatically in our previous work [15] since three subtasks were sufficiently challenging for the user.

Finally, the modified OpenMATB has 10 difficulty levels. Difficulty level is a global setting that simultaneously influences the degree of reticle drift, the frequency of monitoring events, and the frequency of communications instructions. The 10 levels were designed so that level 1 should be very easy for all users while level 10 should be very hard for all users. Specifics about the difficulty levels, including numerical values of subtask difficulty parameters, are available in our previous paper [15].

C. Study Protocol

The protocol was similar to that used in a previous study that explored different aspects of physiological computing [15]. It was approved by the University of Wyoming Institutional Review Board, and all participants gave written informed consent. Overall, each participant was given their assigned sensor (eye tracker or EEG) and interacted with the modified OpenMATB for three 11-minute intervals, with each interval having a higher accuracy of difficulty adaptation. While some elements of the protocol (e.g., use of three intervals) were not needed to answer our research question, they were reused for easier comparison to previous work [15] (which, e.g., also used three intervals). The study involved extensive deception to induce consistently accurate difficulty adaptation; we thus first present the protocol as told to the participant, then the deception aspects.

1) Protocol as Presented to Participant

Upon arrival, participants were told that the study purpose was to test a system that would adapt the difficulty of the modified OpenMATB based on physiological measurements. Participants were told that 3 different algorithms had been developed to extract information from the eye tracker or EEG system (depending on the assigned group) and that they would test all 3 algorithms. They were further told that their difficulty preferences would be collected periodically but would not influence difficulty adaptation; they would only be used to post-hoc verify algorithm performance after the session.

In the EEG group, participants were fitted with an electrode cap. The fitting process was fixed to be 10 min in duration (see Deception subsection) and involved measuring head size with a cloth tape, applying a water-drenched electrode cap, adjusting the cap to obtain a proper fit, and having the participant sit silently while the experimenter prepared recording software on a separate computer. The eye tracker group, by contrast, did not experience any specific sensor setup as the sensor was positioned below the screen prior to the participant's arrival.

The OpenMATB was started and participants practiced with it for 5 min at difficulty level 5 of 10. The experimenter offered advice and answered questions during practice. In the eye tracker group, the eye tracker was then "calibrated" by having participants look at several points on the screen for a few seconds each (~30 s total). In the EEG group, the EEG was "calibrated" by having participants briefly close their eyes, then blink rapidly, then move their eyes in multiple directions, then count backwards silently from 1000 (5 min total).

Participants then interacted with the OpenMATB scenario for three 11-minute intervals, which they were told corresponded to the three algorithms developed for physiological data analysis. In each interval, the OpenMATB was initialized at difficulty level 5. The system then paused every 60 s with a popup text asking participants how they would prefer difficulty to change (options: increase, decrease, no change). After participants made a selection, the OpenMATB resumed at a different difficulty level; participants were not explicitly told how difficulty had changed. At the end of each 11-min interval, participants completed two questionnaires: the Intrinsic Motivation Inventory (IMI – 8-item version as our previous study [15]) and the NASA Task Load Index (TLX) [18]. After the third 11-minute interval, participants were asked how much they liked each of the three algorithms and rated all three on a visual analog scale from "did not like at all" to "liked very much". Participants rated all three simultaneously and were encouraged to consider them relative to each other. This questionnaire was reused from our prior work [14], [15].

2) Deception

Though participants were told that their preferences would not influence difficulty adaptation, this was not true. In reality, the physiological data were not collected at all, and neither sensor was truly calibrated – participants experienced a realistic approximation of a setup and calibration procedure whose duration was fixed. For example, in the EEG group, setup duration was fixed by giving the experimenter a hidden 10-min timer and having them finish "setup" once 10 min had passed. Without physiological data, difficulty adaptation was done based on participants' preferences. Specifically, the 3 intervals had predefined adaptation accuracies: difficulty adaptation followed the participant's preference 70% of the time in the first interval, 80% in the second, and 90% in the third. Since participants were queried for their preference 10 times (every 60 s), 70% agreement was for example induced by following the preference after 7 of 10 queries. When following the participant's preference, difficulty was adapted in the preferred direction by 1 difficulty level (if participant was assigned to the 'small' group) or by 3 levels (if assigned to 'large' group). When not following the preference, difficulty was adapted by 1 or 3 levels in the opposite of the preferred direction (if participant preferred increasing/decreasing difficulty) or a random direction (if they preferred not changing difficulty). Both 'small' and 'large' groups were included since our previous study found significant differences in user experience as a result of this factor [15].

Similar deception without time-consuming sensor setup was used in our previous study [15] and allowed the behavior of the physiological computing system to be kept consistent between participants. A realistic physiological computing

system would have variations in pattern recognition accuracy between participants, and the EEG and eye tracker would have different average accuracies. Furthermore, in a realistic physiological computing system, the setup time for EEG may vary significantly between participants due to factors such as hair. Deception allowed us to keep these aspects of the protocol consistent.

D. Data Analysis

Each participant's self-report data consisted of 6 outcomes for each of the three intervals. The IMI has 4 subscales: interest/enjoyment, effort/importance, perceived competence, and pressure tension; each of these 4 was analyzed as a separate outcome. The NASA TLX has six subscales, which were merged into one TLX score by summing all subscales with performance reversed [18]. Finally, answers to the visual analog scale at the end of the session were converted to a 0-100 numerical scale, dubbed "relative liking". As mentioned, physiological data were not truly collected and were used only as part of the deception, so no physiological data analysis was done.

Each outcome was analyzed with a mixed analysis of variance with one within-subjects factor (interval: 1, 2, 3) and two between-subjects factors (sensor: eye tracker / EEG, adaptation magnitude: small/large). Greenhouse-Geisser corrections were used, and effects are reported as significance (p) and effect size (partial eta squared – η_P^2). For our research question, we were mainly interested in effects of the sensor: a main or interaction effect of the sensor would indicate that the eye tracker led to a different user experience than the EEG.

III. Results

Table I shows main effects of interval, sensor and adaptation magnitude on all six outcomes while Table II shows two-way interaction effects. There was only one effect of sensor with $p < 0.05$: the interval * sensor interaction effect on effort/importance with $p = 0.04$ and $\eta_P^2 = 0.049$. Specifically, effort/importance scores for intervals 1, 2 and 3 were as follows in the two sensor groups:

- EEG: 11.8 ± 1.8, 12.6 ± 1.7, 12.0 ± 2.1
- Eye tracker: 12.2 ± 1.6, 12.2 ± 2.0, 12.4 ± 1.8

IV. Discussion

Main effects of interval were observed on effort/importance and competence: both increased as adaptation accuracy increased. While this does not answer our research question, it shows that better adaptation makes participants apply more effort and feel more competent – a phenomenon also observed in our previous work [15]. The significant main effect of magnitude on interest/enjoyment indicates that enjoyment was higher with large adaptation actions, which is likely because large adaptation actions allowed users to reach a suitable target difficulty more quickly. This is a slightly different result from our previous study [15], which found effects of magnitude on other IMI subscales but not interest/enjoyment; however, we do not discuss this (or the interval * magnitude interaction effects) in further detail since it does not relate to the research question.

TABLE I. Main effects of interval, sensor, and adaptation magnitude on all 6 outcome variables, presented as p-values and partial eta squared. Effects with $p < 0.05$ are bolded.

Outcome	Main effects					
	Interval		*Sensor*		*Magnitude*	
	p	η_P^2	p	η_P^2	p	η_P^2
Relative liking	.37	.015	.73	.002	.91	.000
Task Load Index	.13	.032	.53	.006	.16	.030
Interest/Enjoyment	.96	.001	.49	.036	**.036**	.066
Effort/Importance	**.026**	.055	.46	.009	.087	.044
Competence	**.001**	.16	.83	.001	.76	.001
Pressure/Tension	.44	.012	.72	.002	.064	.052

TABLE II. Two-way interaction effects of interval, sensor, and adaptation magnitude on all 6 outcome variables, presented as p-values and partial eta squared. Effects with $p < 0.05$ are bolded.

Outcome	Interaction effects					
	*Interval * Sensor*		*Interval * Magnitude*		*Sensor * Magnitude*	
	p	η_P^2	p	η_P^2	p	η_P^2
Relative liking	.38	.015	.47	.012	.46	.009
Task Load Index	.73	.005	**.03**	.054	.45	.009
Interest/Enjoyment	.48	.011	.33	.017	.40	.011
Effort/Importance	**.04**	.049	**.016**	.063	.33	.014
Competence	.81	.003	**.003**	.087	.65	.003
Pressure/Tension	.80	.003	.21	.024	.83	.001

With regard to our research question, there is little evidence that different sensors resulted in a different self-reported user experience. While there was one interval * sensor interaction effect, it was small by effect size standards [19] and does not have a clear interpretation. Thus, although the two sensors have very different setup/calibration times (30 s for eye tracker, 15 min for EEG) and discomfort/invasiveness levels (no physical contact for eye tracker vs. wet cap for EEG), this does not seem to be reflected in self-reported opinions of the session.

This surprising result can likely be explained by weaknesses of the study protocol. Each participant only experienced one of the two sensors, and the hardware was thus not a major factor to consider when rating user experience. Previous studies indicate that users are poor at rating physiological computing technologies without prior experience with other technologies [14], [15]; while those studies did not vary the sensor type, a similar finding would likely apply here. In the future, we will thus explore a different protocol to better answer our research question. Specifically, we will have each participant experience two or more sensor types, and we will ask questions to explicitly compare these sensor types.

Nonetheless, results of our study have value for developers and evaluators of similar technologies. They show that, in the absence of experience with more comfortable technologies, participant ratings of technology performance and user experience are overall not affected by discomfort or long setup time. While such issues should still be addressed before such technologies are broadly deployed, they do not appear to negatively impact single-session lab-based evaluations done with participants drawn from the general population.

V. Conclusion

Our study found few differences in self-reported user experience between two physiological computing systems with very different setup times and user comfort levels. Only one effect of sensor type was found, and it was small. This lack of differences is surprising and can likely be explained by the fact that each participant only experienced one sensor and thus likely did not consider this aspect when rating their experience. Nonetheless, it suggests that, in the absence of experience with more user-friendly devices, users' perception of device performance is not influenced by discomfort or long setup time. These factors thus do not negatively impact short evaluations of prototype technologies. Nonetheless, in the future, we will explore an alternative study protocol where each participant experiences two or more sensor types.

References

[1] S. H. Fairclough, "Fundamentals of physiological computing," *Interact. Comput.*, vol. 21, no. 1–2, pp. 133–145, Jan. 2009, doi: 10.1016/j.intcom.2008.10.011.

[2] D. Novak, M. Mihelj, J. Ziherl, A. Olenšek, and M. Munih, "Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 19, no. 4, pp. 400–410, 2011.

[3] S. D'Mello and A. C. Graesser, "Feeling, thinking, and computing with affect-aware learning technologies," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds., Oxford University Press, 2014, pp. 419–434.

[4] A. Darzi, S. Gaweesh, M. Ahmed, and D. Novak, "Identifying the causes of drivers' negative states using driver characteristics, vehicle kinematics and physiological measurements," *Front. Neurosci.*, vol. 12, p. 568, 2018.

[5] M. D. Van der Zwaag, J. H. Janssen, and J. H. D. M. Westerink, "Directing physiology and mood through music: validation of an affective music player," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 57–68, 2013.

[6] R. V. Aranha, C. G. Correa, and F. L. S. Nunes, "Adapting software with affective computing: a systematic review," *IEEE Trans. Affect. Comput.*, vol. 12, no. 4, pp. 883–899, 2021, doi: 10.1109/TAFFC.2019.2902379.

[7] S. D. Kreibig, "Autonomic nervous system activity in emotion: a review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, Jul. 2010, doi: 10.1016/j.biopsycho.2010.03.010.

[8] A. Al-Nafjan, M. Hosny, Y. Al-Ohali, and A. Al-Wabil, "Review and classification of emotion recognition based on EEG brain-computer interface system research: A systematic review," *Appl. Sci.*, vol. 7, p. 1239, 2017, doi: 10.3390/app7121239.

[9] Y. Zheng, B. Tian, Z. Zhuang, Y. Zhang, and D. Wang, "fNIRS-based adaptive visuomotor task improves sensorimotor cortical activation," *J. Neural Eng.*, vol. 19, p. 046023, 2022.

[10] W. Boucsein, *Electrodermal Activity*, 2nd ed. Springer, 2012.

[11] P. Ledwidge, J. Foust, and A. Ramsey, "Recommendations for developing an EEG laboratory at a primarily undergraduate institution," *J. Undergrad. Neurosci. Educ.*, vol. 17, no. 1, pp. A10–A19, 2018.

[12] S. H. Fairclough, A. J. Karran, and K. Gilleade, "Classification accuracy from the perspective of the user: real-time interaction with physiological computing," in *Proceedings of the 33rd Annual Conference on Human Factors in Computing Systems (CHI '15)*, 2015, pp. 3029–3038.

[13] S. H. Fairclough and F. Lotte, "Grand challenges in neurotechnology and system neuroergonomics," *Front. Neuroergonomics*, vol. 1, p. 602504, 2020, doi: 10.3389/fnrgo.2020.602504.

[14] S. M. McCrea, G. Geršak, and D. Novak, "Absolute and relative user perception of classification accuracy in an affective videogame," *Interact. Comput.*, vol. 29, no. 2, pp. 271–286, 2017.

[15] V. D. Novak, D. Hass, M. S. Hossain, A. F. Sowers, and J. D. Clapp, "Effects of adaptation accuracy and magnitude in affect-aware difficulty adaptation for the Multi-Attribute Task Battery," *Int. J. Hum. Comput. Stud.*, vol. 183, p. 103180, 2024, doi: 10.1016/j.ijhcs.2023.103180.

[16] P. Sawangjai, S. Hompoonsup, P. Leelaarporn, S. Kongwudhikunakorn, and T. Wilaiprasitporn, "Consumer grade EEG measuring sensors as research tools: a review," *IEEE Sens. J.*, vol. 20, no. 8, pp. 3996–4024, 2020.

[17] J. Cegarra, B. Valéry, E. Avril, C. Calmettes, and J. Navarro, "OpenMATB: A Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability," *Behav. Res. Methods*, vol. 52, pp. 1980–1990, 2020.

[18] S. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., Amsterdam: North Holland Press, 1988.

[19] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.