Multimodal Approach to Assess a Virtual Reality-based Surgical Training Platform

Doga Demirel [1, *] [0000-0002-8270-1163], Hasan Onur Keles [2] [0000-0001-8493-2582], Chinmoy Modak [1], Kubranur Kara Basturk [2], Jacob R. Barker [1], Tansel Halic [3] [0000-0002-2558-4001]

[1] Florida Polytechnic University, Lakeland, Florida, USA
[2] Ankara University, Ankara, Turkey
[3] Intuitive Surgical, Peachtree Corners, Georgia, USA

*Corresponding Author: ddemirel@floridapoly.edu
Email addresses in order: ddemirel@floridapoly.edu, hokeles@ankara.edu.tr, cmodak9952@floridapoly.edu, kubranur.kara@tootech.com.tr, jbarker@floridapoly.edu, tansel.halic@intusurg.com

Abstract. Virtual reality (VR) can bring numerous benefits to the learning process. Combining a VR environment with physiological sensors can be beneficial in skill assessment. We aim to investigate trainees' physiological (ECG) and behavioral differences during the virtual reality-based surgical training environment. Our finding showed a significant association between the VR-Score and all participants' total NASA-TLX workload score. The extent of the NASA-TLX workload score was negatively correlated with VR-Score (R^2 =0.15, $P < 0.03$). In time-domain ECG analysis, we found that RMSSD (R^2 =0.16, $P < 0.05$) and pNN50 (R^2 =0.15, $P < 0.05$) scores correlated with significantly higher VR-score of all participants. In this study, we used SVM (linear kernel) and Logistic Regression classification techniques to classify the participants as gamers and non-gamers using data from VR headsets. Both SVM and Logistic Regression accurately classified the participants as gamers and non-gamers with 83% accuracy. For both SVM and Linear Regression, precision was noted as 88%, recall as 83%, and f1-score as 83%. There is increasing interest in characterizing trainees' physiological and behavioral activity profiles in a VR environment, aiming to develop better training and assessment methodologies.

Keywords: Virtual Reality, Skill assessment, ECG, Mental workload

1 Introduction

Virtual Reality (VR) is becoming a more widely used teaching and learning aid in several fields, such as medical training and robotics. Conventional training methods are hard to grasp, non-reusable, non-repeatable, and costly. VR allows students and teachers to interact in a real-time learning environment, which would be nearly impossible to do in the physical world. Trainees face trouble obtaining skills in this unnatural environment. Training with VR simulators established benefits, however, methods for skill assessment in VR, particularly in real-time, are still undeveloped and unknown. Virtual reality simulators are computer-based systems that generate output data, which is very helpful for skill assessment [1, 2].

In general, skill assessment approaches can be found in technical, non-technical, and mental workload assessments. For mental workload assessment, questionnaires and physiological measurements can be useful tools, and for non-technical skill assessment, all methods (questionnaires, expert-rating and physiological measurement) can be utilized [3]. In critical fields such as surgical education and healthcare, learning is based on an apprenticeship model [4]. In this model, the proficiency assessment is the responsibility of the trainers. However, their assessment is subjective. Objective assessment is essential because performance in training and performance are difficult to correct without objective feedback. Psychophysiological measures allow a more objective assessment and can provide an uninterrupted evaluation [5].

Technological advances in wearable sensor technology make objective assessment less intrusive and capable of delivering continuous, multimodal information. Electroencephalogram (EEG) and Electrocardiogram (ECG), including Heart Rate (HR), Heart Rate Variability (HRV), have also been correlated with NASA-TLX scores as well as task complexity, performance and expertise in surgery [5, 6]. Studies also indicate that such descriptors correlate with the overt performance of human operators [7]. For example, mental workload gauged by a standard self-reporting tool was proportional to the rate of errors committed and suture quality in laparoscopic surgery training [8]. However, efforts to characterize mental status descriptors and their effect on individual and team performance face a significant challenge: the descriptors are not directly observable. To quantify them, researchers traditionally resort to physiological variables (e.g., ECG, EEG, skin conductance), behavioral indicators (e.g., secondary task performance), or survey results (e.g., NASA-TLX questionnaire). Few studies focused on combining VR environments and physiological sensors during training approach [9, 10, 11, 12].

Previous gaming experience helps get accommodated to this training environment faster, increases visual attention capacity, and makes multitasking easier. Thus, it helps the trainees to facilitate these obstacles and acquire skills more quickly. Video gamers and surgeons have similarities in skill acquisition [13, 14]. Video gamers have superior eye-hand coordination, faster reaction times, superior spatial visualization skills, and a high capacity for visual attention and spatial distribution. Both laparoscopic surgery and computer games require eye-hand coordination, visuospatial cognitive ability, attention, and perception skills. Individuals who interact or play video games tend to have better visuospatial ability when compared to non-gamers [15,16].

Grantcharov et al. [17] demonstrated the effect of video game experience on the MIST-VR® surgical simulator and found that surgeons with previous video game experience made significantly fewer errors than non-gamers. Therefore, this project aims to assess how the gaming experience gives advantages to the new trainees and makes the learning process more accessible on the VR-based surgical training platforms. Demand for the safety of patients has prompted the need for efficient and affordable training for preparing surgeons. Several VR-based simulators have recently been developed to fulfill this need, and VR applications, simulation, and e-learning have improved the learning metrics [18]. Conventional human and animal models, cadavers, and mannequin-based training for surgeons can be risky, non-reusable, subjective, and expensive. VR-based simulators measure several characteristics or metrics for objectively assessing the trainee's performance.

According to Enochsson et al. [19], video game players were more efficient and faster than non-gamers performing the simulated colonoscopy. There were also no gender-specific differences in performance. Jalink et al. [20] suggested that video games could be used to train surgical residents in laparoscopic skills. Based on these findings, one might expect that gaming will facilitate and improve the training of novice surgeons where the performance requires a firm reliance on spatial orientation and the recognition of various visual inputs.

We aim to investigate trainees' physiological and behavioral differences during the virtual reality-based surgical training environment. This paper shows multimodal information collected from a VR+ECG system for skill assessment during a surgical training game. We hypothesize that multimodal information can lead to a more accurate assessment than single modality-based measurement approaches. In addition, we also showed how the game experience could affect performance and behavioral measures (task load). To our knowledge, the studies to date which investigated these links exclusively utilized overt performance and behavioral measures. However, given the complexity of skill assessment, multimodal approaches are required.

2 Methods

2.1 Participants

Our dataset consisted of 30 participants with varying levels of gaming experience (from 0 to 60 hours per week). It was subsequently divided into two groups according to their previous gaming experience as gamers and non-gamers. The Research Ethics Board of Ankara University approved this study (2021/435), which was performed in agreement with the Declaration of Helsinki. All participants signed informed consent and could withdraw from the study at any time.

2.2 Study Design

For this experiment, we created a simple VR environment where the users were asked to bounce a balloon and keep it in a proper range while in the air. To achieve a high score in this environment, the trainee must keep the balloon between the planes while bouncing for as long as possible with fewer impacts while causing no damage to the balloon. A VR score is calculated by the amount of time between the planes. This environment assesses the gentleness of a surgeon [21]. We used a Meta Quest 2 as the VR headset.

The participants were given two minutes to get accommodated with the scene (resting state), and then the next three minutes were used to capture data. In this study, we recorded physiological signals such as ECG while using a VR device. After subjects completed the training in VR headsets, they completed the NASA task load index (NASA-TLX) questionnaire. Nasa-TLX is a multidimensional rating scale that provides an overall index of mental workload and the relative contributions of six subscales: mental, physical, and temporal task demands: effort, frustration, and perceived performance. VR-Score, jerk, velocity, and acceleration were recorded for all participants using a VR headset. RMSSD, pNN50, and pNN20 were used for heart rate variability (HRV) analysis.

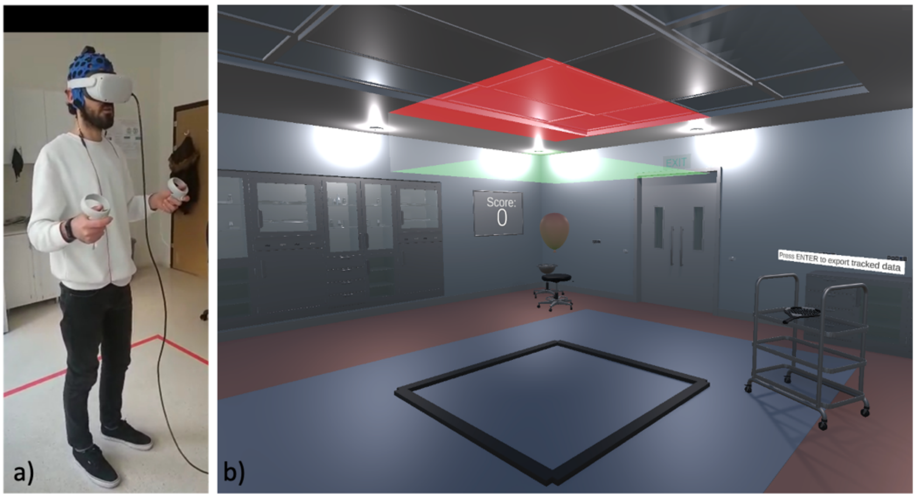

Fig. 1. (a) VR-ECG Setup and (b) VR Racket Game

2.3 Physiological signal recording and processing

The ECG data were obtained through the ExG Explorer device (wearable wireless) (Mentalab, Germany). One channel was recorded by placing the electrodes on the designated body locations. Raw ECG signals were digitalized with a sampling rate of 250 Hz and filtered by a low pass Gaussian filter with a cut-off frequency of 40 Hz, while IIR Zero-Phase Filter was used to attenuate baseline wander with a cutoff frequency of 0.5 Hz. Time domain analysis was used to process the preprocessed ECG signals. The

HRV time domain parameters (RMSSD, pNN 50, and pNN 20) were chosen in the current study as the assumed assessment indicators for the later analysis.

2.4 Statistical Analysis

When conducting a regression analysis comparing two numerical variables, linear fit with analysis of variance was used. The descriptive results comparing two groups, NASA-TLX, and total VR score, contained non-paired data. To assess the statistical significance of the difference between two groups of non-paired results, we used the non-parametric Kolmogorov test. We did not utilize null hypotheses whose rejection would have required corrections for multiple comparisons or false discovery. The statistical significance of the results were interpreted based on p values, and $p < 0.05$ was set as the level of statistical significance.

3 Result

3.1 VR Results

According to our results, gamers had an average VR score of 3581.86, over two times higher than non-gamers average VR score of 1748.8. The time gamers kept the balloon between the planes (66.36) is almost two times higher than the non-gamers (34.96). Also, gamers popped the balloon times above the top plane (24.21) was significantly higher than the non-gamers (12.8). The mean jerk was 40.9%, the mean acceleration was 29.1%, and the mean velocity was 7% more in non-gamers than gamers. While for mean path length, gamers had 7.1% more than non-gamers. For standard deviation results, gamers had a lower standard deviation in path length (7.6%), velocity (30.1%), acceleration (27.8%), and jerk (23.6%). These findings indicate that gamer's hand and spatial movement were gentler, with a minor standard deviation, than non-players.

The results of the features of the compared groups are illustrated in Figure 2. The gamer's hand position (Figure 2a) and path length (Figure 2b) while moving the tennis racket are more scattered than the non-gamers. Figure 2b also shows that gamers had a longer mean path length showing that they were more decisive and knew what they were doing. This finding implies gamer has better hand-eye coordination as they perform better and produce better result than non-gamers. However, from Figure 2c, we can see that non-gamers had more velocity meaning that gamers used more positive force. As gamers have less acceleration (Figure 2d) and jerk (Figure 2e), their movement is more stable than the non-gamers.

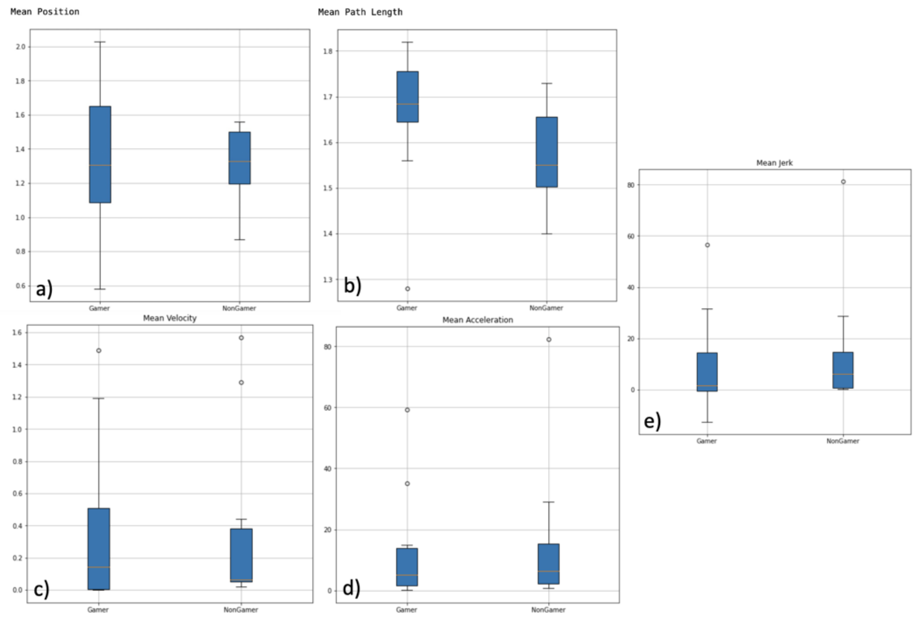

Fig. 2. (a) Mean Position, (b) Mean Path Length, (c) Mean Velocity, (d) Mean Acceleration, and (e) Mean Jerk box plots comparing Gamer and Non-gamer results.

3.2 Clustering Results

After selecting the features from the data set, we used multiple metrics to measure the difference between the data groups. We used the Davies Bouldin score, Silhouette Score, and the Mutual Information Index metrics with the K-Means, Mean Shift, and Spectral Clustering algorithms for clustering. We got the optimum score at the number of clusters (n) = 2. Figure 3 shows the graphs for the metrics score of the clustering algorithms. Normalizing the data, other better results over Spectral Clustering and Mean Shift algorithm. Though there are variations in the results, Mean-Shift performed the best, achieving an 80% success rate in classifying the users based on their previous gaming experience. However, since the range of data is smaller, the Mutual Information Index dropped in almost all instances.

Fig. 3. Clustering Results according to the gaming experience

3.3 Classification Results

After the clustering, we passed the data through classification algorithms such as Logistic Regression and Support Vector Machine with linear kernels. We looked at those classification algorithms' precision, recall, F1 score, and average accuracy. We normalized the data the same way as above and obtained massive improvements in the Logistic regression algorithm. We could classify the performance between the gamers and non-gamers at best 83% of the time.

Table 1. Classification results according to the gaming experience

Algorithms	Precision	Recall	F1 Score
Logistic Regression	80%	67%	62%
SVM Linear	88%	83%	83%
AdaBoost	88%	83%	83%

Table 1 lists the Precision, Recall, and F1 Scores of different classification algorithms. We achieved at best, 88% of the average score for those algorithms. The distribution of the observed values is displayed more clearly in Figure 4. We predicted that the gamers would perform than the participants with no gaming experience, and in the case of all classification algorithms, the True Positive (TP) is significantly higher. Logistic regression successfully classified gamers with 82% accuracy and non-gamers with 86% accuracy (Figure 4a). Logistic regression successfully classified gamers with 82% accuracy and non-gamers with 86% accuracy. Figure 4b shows SVM Linear classified gamers with 91% accuracy and non-gamers with 86% accuracy. Adaboost Classifier classified both gamers and non-gamers with 100% accuracy (Figure 4c)

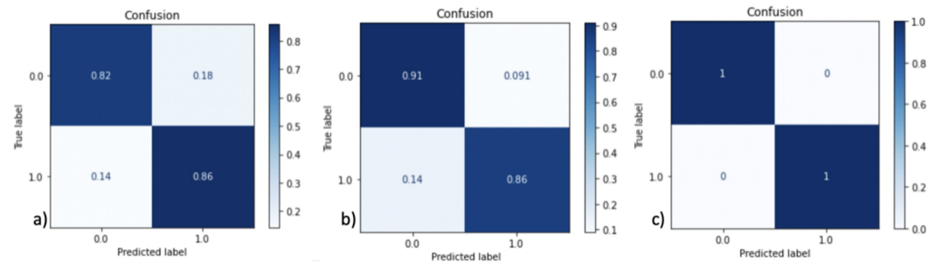

Fig. 4. Confusion matrix for (a) Logistic Regression, (b) SVM Linear Kernel, and (c) AdaBoost classifier

3.4 Behavioral and Physiological Results

We present the behavioral and subjective metrics and physiological measurement (ECG) of non-gamer and gamers, measured while performing virtual training tasks using a VR headset. Thirty participants (mean age 23.25 ± 8.5 years) were enrolled in the study. The participants' experience with gaming varies. Fifteen participants have no experience with gaming. The rest of the participants have an experience with gaming (mean: 12.4 hours per week± 11.4 years).

Figure 5a shows that gamers have greater VR scores when compared to non-gamers. A significant difference in VR Score was noted between gamers and non-gamers ($P < 0.05$), confirming the difference between the two groups. Figure 5b, through regression analysis, we found a significant association between the VR score and Nasa-TLX total score for all participants. The extent of the Nasa-TLX score (MW) was negatively correlated with the VR performance score ($R^2 = 0.14$, $P < 0.05$). Although gamers have a higher Nasa-TLX score, the difference between gamers and non-gamers did not reach significance during the racket game.

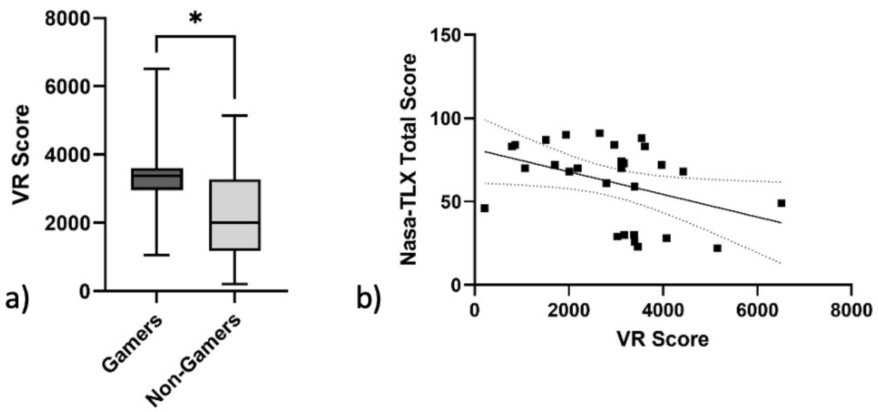

Fig. 5. (a)VR Performance of non-gamer students (gray, circle indicates the median) vs. gamer students (black). Error bars indicate sample standard deviations (*p<0.05; **p< 0.01). (b) Nasa-TLX Total Score vs. VR Score

In time-domain ECG analysis, we found that RMSSD (R^2 =0.16, P < 0.05) and pNN50 (R^2 =0.15, P < 0.05) scores correlated with significantly higher VR-score of all participants (Figure 6). There were similar trends for pNN20, although group differences were not statistically significant.

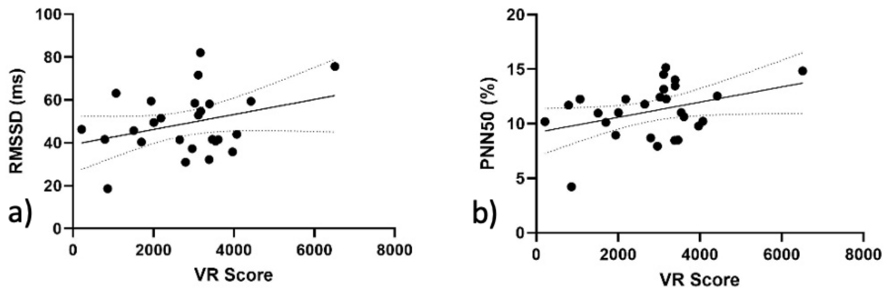

Fig. 6. (a) VR score vs RMSSD (R^2 = 0.14 P<0.05). (b) VR Score vs PNN50 (R^2 = 0.16 P<0.05). In the scatter plots, the solid black lines indicate the linear best fit to the data points, and the dotted lines indicate the 95% confidence interval.

4 Discussion

Our main results suggest a relationship between VR performance scores and the ECG time domain parameters of trainees. To our knowledge, this is the first study that includes physiological and performance metrics (multimodal approach) during a VR environment using a VR headset. Time-domain approach for ECG has been widely used to investigate the cardiovascular outcome of mental work. Several studies showed that mental workload leads to a decrease in the time domain measure of ECG [22, 23]. This supposes a predominant increase in sympathetic activity or a predominant decrease in para sympathetic activity [24, 25].

Our results showed a significant negative correlation between cognitive load and gaming experience in all participants. This outcome is important because previous research has demonstrated that scores from NASA-TLX can accurately predict future performance [6, 26]. Furthermore, playing VR games that have high levels of cognitive demand may result in being easily distracted, having limited options to consider, or being rigid in selecting strategies. On the other hand, a low load allows greater amounts of data to be processed, leading to appropriate responses to unexpected events [27].

There were some key limitations to our study which should be mentioned. Our study had a low number of participants. A higher number of participants would allow us to show the significance of some of the trends observed. The measurement of video game experience may not be entirely accurate due to self-reporting approach.

5 Conclusion

There is increasing interest in characterizing trainees' physiological and behavioral activity profiles in a VR environment, aiming to develop better training and assessment methodologies. We evaluated the benefits of gaming for the new trainees and how it improves learning accessibility on VR-based surgical gentleness training platforms. We conducted investigations involving human subjects primarily to establish content and construct validations. We used different kinematics data, such as position, path length, acceleration, jerk, and velocity, collected from the subject's interactions with the virtual reality environment. The dataset for this project consisted of 30 participants, who were then divided into groups based on their prior gaming experience. We depicted that gamer's hand and spatial movement were gentler, with a minor standard deviation, than non-players. Then, we distinguished between gamers and non-gamers by utilizing a variety of clustering and classification algorithms. We applied clustering algorithms such as K-means, Mean-shift, and Spectral Clustering to verify the difference between the data. Though there are variations in the results, Mean-Shift performed the best, achieving over an 80% success rate in classifying the users based on their previous gaming experience. Using Logistic Regression, Support Vector Machine (SVM), and AdaBoost classifier, we were able to classify over 80% of the performance between the gamers and non-gamers and achieved 88% accuracy at best. Gamer's subjective performance and situation awareness correlated more positively with task performance than the non-gamers.

Acknowledgement

This project was made possible by NIH/NIAMS 5R44AR075481- 03 and the Arkansas INBRE program, supported by the National Institute of General Medical Sciences (NIGMS), P20 GM103429 from the National Institutes of Health (NIH), and partially supported by NIH/NIBIB 1R01EB033674-01A1, 5R01EB025241-04, 3R01EB005807-09A1S1, and 5R01EB005807-10.

References

1. Yu, P., Pan, J., Wang, Z., Shen, Y., Li, J., Hao, A., & Wang, H. (2022). Quantitative influence and performance analysis of virtual reality laparoscopic surgical training system. *BMC medical education*, *22*(1), 1-10.

2. Ota D, Loftin B, Saito T, Lea R, Keller J. Virtual reality in surgical education. Comput Biol Med. 1995 Mar;25(2):127-37. doi: 10.1016/0010-4825(94)00009-f. PMID: 7554831.
3. Dias RD, Ngo-Howard MC, Boskovski MT, Zenati MA, Yule SJ. Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. Br J Surg. 2018 Apr;105(5):491-501. doi: 10.1002/bjs.10795. Epub 2018 Feb 21. PMID: 29465749; PMCID: PMC5878696.
4. Beard JD. Assessment of surgical skills of trainees in the UK. Ann R Coll Surg Engl. 2008 May;90(4):282-5. doi: 10.1308/003588408X286017. PMID: 18492389; PMCID: PMC2647187.
5. Zakeri, Z., Mansfield, N., Sunderland, C. et al. Physiological correlates of cognitive load in laparoscopic surgery. Sci Rep 10, 12927 (2020). https://doi.org/10.1038/s41598-020-69553-3
6. Keleş, H. O., Cengiz, C., Demiral, I., Ozmen, M. M., & Omurtag, A., (2021). High density optical neuroimaging predicts surgeons's subjective experience and skill levels. PLOS ONE , vol.16, no.2.
7. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Med Educ. 2010 Jan;44(1):85-93. doi: 10.1111/j.1365-2923.2009.03498.x. PMID: 20078759.
8. Yurko, Yuliya Y. MD; Scerbo, Mark W. PhD; Prabhu, Ajita S. MD; Acker, Christina E. BA; Stefanidis, Dimitrios MD, PhD. Higher Mental Workload is Associated With Poorer Laparoscopic Performance as Measured by the NASA-TLX Tool. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare 5(5):p 267-271, October 2010.
9. Suárez, J. X., Gramann, K., Ochoa, J. F., Toro, J. P., Mejía, A. M., & Hernández, A. M. (2022). Changes in brain activity of trainees during laparoscopic surgical virtual training assessed with electroencephalography. Brain Research, 1783, 147836.
10. Yu, P., Pan, J., Wang, Z., Shen, Y., Wang, L., Li, J., Hao, A. and Wang, H., 2021, March. Cognitive load/flow and performance in virtual reality simulation training of laparoscopic surgery. In 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 466-467). IEEE.
11. Tremmel C, Herff C, Sato T, Rechowicz K, Yamani Y, Krusienski DJ. Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG. Front Hum Neurosci. 2019 Nov 14;13:401. doi: 10.3389/fnhum.2019.00401. PMID: 31803035; PMCID: PMC6868478.
12. Aspiotis V, Miltiadous A, Kalafatakis K, Tzimourta KD, Giannakeas N, Tsipouras MG, Peschos D, Glavas E, Tzallas AT. Assessing Electroencephalography as a Stress Indicator: A VR High-Altitude Scenario Monitored through EEG and ECG. Sensors (Basel). 2022 Aug 3;22(15):5792. doi: 10.3390/s22155792. PMID: 35957348; PMCID: PMC9371026.
13. Nahum M, Bavelier D (2020) Video games as rich environments to foster brain plasticity. Handb Clin Neurol 168:117–136
14. Rosser JC Jr, Lynch PJ, Cuddihy L, Gentile DA, Klonsky J, Merrell R (2007) The impact of video games on training surgeons in the 21st century. Arch Surg 142(2):181–186; discussion 6
15. Datta R, Chon SH, Dratsch T, Timmermann F, Muller L, Plum PS et al (2020) Are gamers better laparoscopic surgeons? Impact of gaming skills on laparoscopic performance in "Generation Y" students. PLoS One 15(8):e0232341
16. Chalhoub E, Tanos V, Campo R, Kesrouani A, El Rassy E, Rizkallah J, Chalhoub D, Walter C, Sleiman Z (2016) The role of video games in facilitating the psychomotor skills training in laparoscopic surgery. Gynecol Surg 13:419–424

17. Grantcharov, T.P., Bardram, L., Funch-Jensen, P. and Rosenberg, J., 2003. Impact of hand dominance, gender, and experience with computer games on performance in virtual reality laparoscopy. Surgical Endoscopy and Other Interventional Techniques, 17, pp.1082-1085.
18. Thijssen, A.S. and Schijven, M.P., 2010. Contemporary virtual reality laparoscopy simulators: quicksand or solid grounds for assessing surgical trainees?. The American Journal of Surgery, 199(4), pp.529-541.
19. Enochsson, L., Isaksson, B., Tour, R., Kjellin, A., Hedman, L., Wredmark, T. and Tsai-Felländer, L., 2004. Visuospatial skills and computer game experience influence the performance of virtual endoscopy. Journal of gastrointestinal surgery, 8, pp.874-880.
20. Jalink, M.B., Goris, J., Heineman, E., Pierie, J.P.E. and Henk, O., 2014. The effects of video games on laparoscopic simulator skills. The American Journal of Surgery, 208(1), pp.151-156.
21. Farmer, J., Demirel, D., Erol, R., Ahmadi, D., Halic, T., Kockara, S., Arikatla, V.S., Sexton, K. and Ahmadi, S., 2020. Systematic approach for content and construct validation: case studies for arthroscopy and laparoscopy. The International Journal of Medical Robotics and Computer Assisted Surgery, 16(4), p.e2105.
22. Fan X, Zhao C, Zhang X, Luo H, Zhang W. Assessment of mental workload based on multi-physiological signals. Technol Health Care. 2020;28(S1):67-80. doi: 10.3233/THC-209008. PMID: 32364145; PMCID: PMC7369076.
23. Yu, D., Antonik, C.W., Webber, F. et al. Correction to: Multi-modal physiological sensing approach for distinguishing high workload events in remotely piloted aircraft simulation. Hum.-Intell. Syst. Integr. 3, 201–211 (2021). https://doi.org/10.1007/s42454-021-00033-3
24. Weippert M, Behrens K, Rieger A, Stoll R, Kreuzfeld S. Heart rate variability and blood pressure during dynamic and static exercise at similar heart rate levels. PLoS One. 2013 Dec 13;8(12):e83690. doi: 10.1371/journal.pone.0083690. PMID: 24349546; PMCID: PMC3862773.
25. McCraty R, Shaffer F. Heart Rate Variability: New Perspectives on Physiological Mechanisms, Assessment of Self-regulatory Capacity, and Health risk. Glob Adv Health Med. 2015 Jan;4(1):46-61. doi: 10.7453/gahmj.2014.073. PMID: 25694852; PMCID: PMC4311559
26. Adapa K, Pillai M, Das S, Mosaly P, Mazur L. Predicting Objective Performance Using Perceived Cognitive Workload Data in Healthcare Professionals: A Machine Learning Study. Stud Health Technol Inform. 2022 Jun 6;290:809-813. doi: 10.3233/SHTI220191. PMID: 35673130.
27. Carswell CM, Clarke D, Seales WB (2005) Assessing mental workload during laparoscopic surgery. Surg Innov 12(1):80–90