HHS Public Access

Author manuscript

Appl Ergon. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Appl Ergon. 2021 January ; 90: 103251. doi:10.1016/j.apergo.2020.103251.

Sensor-based indicators of performance changes between sessions during robotic surgery training

Chuhao Wu,
Purdue University, West Lafayette, Indiana, United States

Jackie Cha,
Purdue University, West Lafayette, Indiana, United States

Jay Sulek,
Indiana University, Indianapolis, Indiana, United States

Chandru P. Sundaram,
Indiana University, Indianapolis, Indiana, United States

Juan Wachs,
Purdue University, West Lafayette, Indiana, United States

Robert W. Proctor,
Purdue University, West Lafayette, Indiana, United States

Denny Yu
Purdue University, West Lafayette, Indiana, United States

Abstract

Training of surgeons is essential for safe and effective usage of robotic surgery, yet current assessment tools for learning progression are limited. The objective of this study was to measure changes in trainees' cognitive and behavioral states as they progressed in a robotic surgeon training curriculum at a medical institution. Seven surgical trainees in urology who had no formal robotic training experience participated in the simulation curriculum. They performed 12 robotic skills exercises with varying levels of difficulty repetitively in separate sessions. EEG (electroencephalogram) activity and eye movements were measured throughout to calculate three metrics: engagement index (indicator of task engagement), pupil diameter (indicator of mental workload) and gaze entropy (indicator of randomness in gaze pattern). Performance scores (completion of task goals) and mental workload ratings (NASA-Task Load Index) were collected after each exercise. Changes in performance scores between training sessions were calculated. Analysis of variance, repeated measures correlation, and machine learning classification were used

Corresponding author: Denny Yu, dennyyu@purdue.edu, 315 N. Grant Street, Grissom Hall Room 268, West Lafayette, IN 47907.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

to diagnose how cognitive and behavioral states associate with performance increases or decreases between sessions. The changes in performance were correlated with changes in engagement index (r_{rm} = −.25, p < .001) and gaze entropy (r_{rm} = −.37, p < .001). Changes in cognitive and behavioral states were able to predict training outcomes with 72.5% accuracy. Findings suggest that cognitive and behavioral metrics correlate with changes in performance between sessions. These measures can complement current feedback tools used by medical educators and learners for skills assessment in robotic surgery training.

Keywords

Robotic Surgery; Eye tracking; electroencephalogram; Simulated Training; Performance

1. Introduction

Robotic-assisted surgery (RAS) enables surgeons to tele-operate on patients using instruments inserted through small incisions (Mack, 2001; Sackier & Wang, 1994). This technique minimizes the invasiveness of the procedure to patients and can potentially reduce blood loss, shorten postoperative stay, and result in other patient benefits (Diana & Marescaux, 2015; Giulianotti et al., 2011; Mack, 2001). In addition, RAS can augment surgeon performance by enabling increased dexterity, more ergonomic body positions, and improved visualization (Lanfranco et al., 2004).

Despite these potential benefits, the novel technology in RAS operations provide unique challenges to trainees learning the technique. For example, unlike conventional open surgery, anatomical structures are viewed on a video display under magnification, and surgeons need to adjust themselves to searching and processing visual information from this perspective (Rassweiler et al., 2001). The complex eye-hand coordination, bimanual coordination, and active foot coordination (Da Vinci S System User Manual, 2014; Narazaki et al., 2006) required in RAS may also induce high mental workload for surgeons. Extensive effort has been made to develop computer-based skills simulation approach to RAS training (Morris, 2005). A key advantage of computer-based simulation is that the well-designed tasks allow automatic calculation of objective performance scores by measuring completion time, economy of motion, and number of errors (Brinkman et al., 2013; Lerner et al., 2010). While these scores are typically the only mechanism for evaluating learning progression, there are several limitations. First, performance scores summarize behaviors throughout the task and may not provide real-time information for specific moments that are challenging for the trainee. More importantly, simulators calculate the score based on the task goals, which are mostly related to manual movement. However, cognitive load and visual behaviors can be equally significant for surgeons in live surgeries (Yu et al. 2016) and should be considered during RAS training.

For example, even though a trainee’s performance is no longer improving, decreasing mental workload can be viewed as evidence for learning and improvement (Ruiz-Rabelo et al., 2015). Operators have limited cognitive capacities/resources, and these resources are needed to meet the demands of ongoing tasks (Carswell et al., 2005). Thus, mental workload can

provide insights into where task demands exceed capacity and may hamper performance (Cain, 2007; Wickens, 2008). Studies have observed that more experienced operators report lower mental workload than novices in the same task (Hu et al., 2016; Patten et al., 2006). In addition, many studies have shown that task engagement impacts learning (Berka et al., 2007; Gardner et al., 2016). Engagement usually refers to participants' efforts to maintain sustained attention (Lelis-Torres et al., 2017; Pope et al., 1995). Task engagement is especially relevant to performance when a task requires a high level of vigilance (Horrey et al., 2017; Matthews et al., 2017), as in RAS. The RAS environment also requires high situation awareness of the surgeons (Lai & Entin, 2005; Schiff et al., 2016). Finally, one of the emerging behavioral assessments in surgical education is surgeons' eye gaze patterns/ visual searching skills. By comparing eye gaze patterns, studies found that more experienced surgeons (experts) tended to have similar gaze patterns that differed from those of novices in the same tasks (Khan et al., 2012; Wilson et al., 2010). Teaching experts' gaze patterns to trainees has been suggested as a possible way to improve trainee performance in laparoscopic tasks and accelerate their learning process (Chetwood et al., 2012, Wilson et al., 2011). Yet, additional evidence is needed to determine the impact and potential applications of eye-gaze patterns in robotic surgery skills training.

In summary, cognitive and behavioral measures (i.e., mental workload, engagement, and gaze patterns) may serve as important indicators of surgical skills. However, research is needed to examine how these measures change during the surgical training. This study uses eye-tracking and EEG sensors to record real-time data for measurements and focuses on changes that occur between training sessions. The within-subjects method of comparing changes between RAS sessions helps to focus on the impact of learning by controlling for individual differences and task differences. Changes in performance are used as hypothesized to reflect learning. For example, an increase in performance score represent the assumption that trainee's skills have improved, while a decrease indicates there is no improvement. Based on this framework, two hypotheses are tested:

> Hypothesis 1 (H1): Changes in the performance score between sessions correlate with corresponding changes in cognitive and behavioral in measurements
>
> Hypothesis 2 (H2): Changes of cognitive and behavioral measurements can be used in machine learning to classify whether the performance score is higher than the last session

2. Materials and methods

2.1. Participants

This study was reviewed by the universities' Institutional Review Boards. The study population was surgical trainees (residents in urology and medical students) with no formal training experience in RAS and were training to learn this technique. Seven surgical trainees from a large academic medical school volunteered to participate. All participants were righthand dominant, four were female, and mean age was 26 ±1.6 years. Participants attended multiple training sessions based on their availability.

2.2. RAS training

RAS training was performed on the dVSS (da Vinci Skill Simulator, Intuitive Surgical, Inc. Sunnyvale, CA) platform, with pre-installed task simulation software. Based on recommendations from the clinicians and literature (Alzahrani et al., 2013; Perrenot et al., 2012), six tasks were selected and repeated for each training session. Tasks focused on the skills camera control, EndoWrist instrument manipulation, clutching, needle control, and needle driving. Task names and descriptions are given in Table 1. Most tasks contained 2-3 levels of difficulty, all of which were used in the training session. A task performed at a certain level will be called an *exercise*. Following previous studies (Finnegan et al., 2012; Kenney et al., 2009; Wu et al., 2019), tasks were performed in the order of Camera Targeting, Peg Board, Ring and Rail, Sponge Suturing, Dots and Needles, and Tubes; and lower levels of task difficulty were performed before higher levels. Participants performed the same six tasks at each training session, performing them as much as possible within the 45-minute session (and additional 15 minutes were allocated for study setup). In total, 26 training sessions were schedule over 3 months. The session intervals for each participant were typically two weeks.

2.3. Measurements

2.3.1. Task Performance—

The software in dVSS had a built-in scoring system that assessed trainees' performance based on several criteria, including completion time, economy of motion, number of drops, instrument collisions, excessive instrument force, instruments out of view, and master workspace range (Da Vinci S System User Manual, 2014). This performance score ranged from 0 to 100, with higher scores representing better performance.

2.3.2. Cognitive and behavioral measurements

Mental workload: Although mental workload is commonly assessed using post-task questionnaires (Hart & Staveland, 1988; Wilson et al., 2011), passive sensing approaches such as heart rate measures are increasingly being used to overcome various limitations of post-task subjective questionnaires (Charles & Nixon, 2019; Guru, Shafiei et al., 2015; Marinescu et al., 2018). For this study, pupil diameter was used to assess mental workload, because it was relatively reliable to collect in the RAS environment, i.e., participants were seated with visual information delivered through a robotic console (Figure 1). Pupil dilation had been observed to increase during cognitive activities (Ambler et al., 1977; Beatty, 1982; Beatty & Kahneman, 1966; Palinko et al., 2010; Pomplun & Sunkara, 2003). Previous studies showed that increased pupil diameter was associated with increased workload in surgery (Zhang et al., 2017; Zheng et al., 2015). In the present study, pupil diameter was calculated by taking the mean of left and right pupil diameters.

Engagement: Electroencephalography (EEG) is a common technique for quantifying engagement (Berka et al., 2007; Guru, Esfahani, et al., 2015). Pope et al. (1995) proposed four candidate EEG metrics for measuring task engagement and concluded that Engagement Index (EI) was the best metric due to its sensitivity to different task demands. This index was used in several studies to assess engagement (Chaouachi & Frasson, 2010; Freeman et

al., 1999, 2004), and is defined as the ratio of three power spectral density frequency bands on the parietal montage:

$$EI = Beta / (Alpha + Theta)$$

The ratio between theta, alpha, and beta was based on assumptions that increases in alpha and theta power are associated with relaxed stated or loss of attention (Foxe & Snyder, 2011; Hermens et al., 2005; Sharma & Singh, 2015; Wascher et al., 2014), whereas higher activation in beta band is accompanied with increased alertness (Gola et al., 2013; Kamiński et al., 2012).

Gaze pattern: Eye gaze pattern was quantified using gaze entropy, a metric that describes the randomness of gaze points' distribution (Di Nocera et al., 2007). It was calculated based on Shannon's (1948) entropy theory:

$$H_g(X) = -\sum p(x, y) \cdot \log_2 p(x, y)$$

where $p(x, y)$ is the probability of gaze falling in the (x, y). A gaze point was estimated as coordinates in relation to the 2-dimensional field of view (1920×1080). Gaze entropy for an exercise was calculated based on all gaze points that were monitored during the exercise, across all possible x and y in the field of view.

2.3.3. Subjective metrics—In addition to physiological metrics, subjective metrics were used to complement the sensing metrics for mental workload and engagement. The NASA Task Load Index (NASA-TLX) was used, which contains six subscales of workload (mental demand, physical demand, temporal demand, performance, effort, and frustration) (Hart & Staveland, 1988). Each dimension was rated on a visual analogue scale that ranged from 0 (very low) to 10 (very high). The effort subscale was analyzed to determine consistency with the engagement measures (Venables & Fairclough, 2009). The summation of all six subscales (Raw-TLX) was also analyzed as an indicator of overall perceived workload (Hart, 2006).

2.4. Equipment

A wearable eye-tracking system, Tobii Pro Glasses 2.0, (Tobii Technology AB, Danderyd, Sweden) was used to binocularly sample eye movement at 50 Hz. The eye-tracking sensor was mounted on the glasses frame and connected to a belt-worn recording unit. The system estimated pupil diameter and gaze points and recorded the field-of-view of the wearer. This wearable device can be easily implemented in training and real surgery environment. The recordings were annotated using the Tobii Pro Lab Software (Tobii Technology AB, Danderyd, Sweden) to extract data for pupil diameter and gaze entropy.

EEG signals were collected using a light-weight wireless EEG device (EMOTIV EPOC) and EMOTIV Pro software. Signals were sampled at 128 Hz on 14 channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4. The device used two reference points: CMS/DRL references at P3/P4 and left/right mastoid process. It has been shown that the

device acquire comparable data quality to other EEG devices (Benitez et al., 2016; Stytsenko et al., 2011). Adapting from Pope et al. (1995), we used posterior channels (P7 and P8) to calculate the engagement index. Previous research suggested that the parietal lobes are important in the control of attention and specifically attentional shifts (Posner, 1988; Posner & Petersen, 1990).

2.5. Study procedure

Each session occurred in the operating room during times when no surgeries were scheduled. After arriving at the operation room, the participants reviewed a study information sheet and completed the demographic questionnaire. They were then fitted with the eye-tracking and EEG devices. The systems were calibrated at the beginning of each session per manufacturer instruction. Based on previous studies (Beatty & Lucero-Wagoner, 2000; Marshall, 2000; Mosaly et al., 2017), we collected baseline pupil diameter by having each participant looked at the center of a white screen for 10 s (minimum diameter) and then a black screen (maximum diameter) for 10 s.

Instructions for basic operations of the console (e.g. functions of buttons, and foot pedals) were provided to all participants in their first session. Although they were allowed to familiarize themselves with the controls, no practice sessions on the study tasks were provided so as to capture data from the beginning of the learning. During each task, the console would display pre-programmed messages on task goals and operations, and a researcher was present to address any questions or concerns throughout the session. In each session, participants were expected to perform 12 exercises. To maintain consistency with the trainees' training exposure, the time allocated for performing the simulation tasks of each session was 45 min, including time to complete the surveys, but excluding device setup and calibration time. Although infrequent, some participants were not able to complete all tasks within the time constraint. After completing each exercise, the participant completed a NASA-TLX survey. Eye-tracking and EEG signals were continuously recorded throughout the entire session.

2.6. Data analysis

2.6.1. Data processing—Pupil diameter and gaze entropy were normalized using the feature scaling formula below (Jayalakshmi & Santhakumaran, 2011) to scale the data to the range of [0,1], accounting for potential variation from individual difference in pupil diameter and facilitating the comparison between different variables. It also prevents distortion in analysis caused by variable magnitude difference (Al Shalabi et al., 2006). The value 0 denotes the minimum value for an individual and 1 denotes the maximum for an individual across all sessions.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

EEG signals were processed with the EEGLAB toolbox in MATLAB (Delorme & Makeig, 2004). Specifically, the signals were first re-referenced to average, which rests on the fact that outward positive and negative currents, summed across an entire sphere will sum to 0.

Then it was filtered using a high-pass filter of 0.5 Hz. CleanLine plugin in EEGLAB (Mullen, 2012) was used to adaptively estimate and remove sinusoidal artifacts. Blinks and eye movements artifacts were automatically removed with Independent Component Analysis by using the ADJUST plugin (Mognon et al., 2011). A Fast Fourier Transform (Singleton, 1979) was then applied to transform the signals into a power spectral density (PSD). Three frequency bands were extracted from the PSD: theta (4-8 Hz), alpha (8-12 Hz), beta (12-25 Hz). Average EEG power in P7 and P8 (parietal) were combined to calculate EI for participants.

2.6.2. Session variations—This study focused on the change of each measurement after practice. Therefore, we calculated metrics changes between sessions in sequence, which were named "session variations." The definitions for each session variation variable are shown in Table 2. The first column gives the metric (and its inferred measurement), the second column gives the symbol of the metric session variation, and the third column gives the definition/calculation for the variation. For example, the variation in performance was defined as:

$$\Delta P_{i,j,k} = P_{i,j,k} - P_{i,j,k-1}$$

Where, $i = 1,2,\cdots,7$ represents the subject ID, $j = 1,2,\cdots,12$ represents the exercise ID and $k = 2,3,4,5$ represents the session ID for individuals. Therefore, $\Delta P_{i,j,k}$ is how much the performance of exercise j changed from session $k-1$ to session k for participant i.

2.6.3. Categorization of performance—In order to examine the hypotheses, the dataset was categorized into two groups (improvement or no improvement) based on ΔP. Observations of $\Delta P \leq 0$ were categorized in the no improvement group. To form equal groups for analysis (further details in statistical analysis section), we sampled an equal number of sessions with the largest ΔP across all participants. Specifically, observations in the middle of the distribution were not included in the improvement group due to smaller performance increases that may not as representative for performance improvements. In order to examine the consistency of the improvement/no improvement categorization based on ΔP, we also used a quantile-based threshold as additional analysis. The 1/3 quantile and 3/3 quantile were randomly selected as thresholds for improvement and no improvement group respectively as a systematic threshold for categorizing the data. Results for the quantile-based categorization were consistent with the $\Delta P \leq 0$ categorization and included in the Appendix.

2.6.4. Statistical analysis—For H1, the analysis of variance (ANOVA) was used to examine the impact of the two improvement groups on the physiological and behavioral measures. To test if the session variations of cognitive/behavioral metrics differed for the two improvement groups, ΔM, ΔE and ΔS were used as response variable for separate ANOVA models respectively. Three subjective metrics: $\Delta NASA$, $\Delta NASA^M$ and $\Delta NASA^E$ were also analyzed using ANOVA to compare results from ΔM and ΔE.

Furthermore, H1 was also assessed by the correlation between changes in performance and changes in physiological metrics using repeated measures correlation: r_{rm} (Bakdash & Marusich, 2017). Compared with the common Pearson correlation, r_{rm} coefficient was estimated using ANCOVA (Analysis of Covariance), where participant was treated as a factor level. This technique accounted for underlying subject effects. The formula of r_{rm} combined two linear regression models. The first model considered the subject effect and correlation effect: $Measure\ 1 = \beta_0 + \beta_1 \times Participants + \beta_2 \times Measure\ 2 + \epsilon$. The second model considered the participant effect only: $Measure\ 1 = \beta_0 + \beta_1 \times Participants + \epsilon$. Residual sum of squares of the first model was expressed as SS_{Error}. Residual sum of squares of the second model was expressed as $SS_{Measure} + SS_{Error}$. The correlation was calculated by:

$$r_{rm} = \sqrt{\frac{SS_{Measure}}{SS_{Measure} + SS_{Error}}}$$

For H2, machine learning algorithms were used to determine the accuracy of using multiple metrics to classify the improvement and no improvement groups. Three different algorithms were used: logistic regression, Naïve Bayes algorithm, and Support Vector Machine (SVM) (James et al., 2013). Other studies have demonstrated these approaches for workload classification (So et al., 2017; Solovey et al., 2014). For SVM, linear kernel specification was used. The k-fold cross validation procedure was used for model training and testing (Hastie et al., 2001), and three folds were performed. The equal number of improvement and no-improvement labels prevented the problem of imbalanced data (Batista et al., 2004; Chawla et al., 2004). A confusion matrix (Fawcett, 2006) was used to determine the accuracy and sensitivity of eye metrics in predicting workload. Significance level was set at $\alpha = 0.05$. Multiple comparisons were corrected using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

3. Results

3.1. Performance variation

Over the study period, a total of 26 sessions (294 exercises) were collected from 7 participants. One participant only attended 1 session and was excluded from the analysis. Four participants attended 3 sessions, 2 participants attended 5 sessions, and 1 participant attended 4 sessions. Figure 2 shows how average performance across participants in each task shifted from session 1 to session 5.

Session variations were calculated for 212 exercises ($k \geq 2$). Among these 212 exercises, 61 observations have ΔP below 0 (non-improvement). We matched the number of observations by sampling the same amount of cases (61) from the highest performance increases observations ($\Delta P \geq 13$) as the improvement category (see Figure 3). Across participants the average time interval between sequential sessions was 21.7 days with standard deviation of 13.6 days. The correlation between ΔP and session time interval was −.37($p < .001$), where longer time intervals result in worse performance change. Thus, time was included as a covariate in the model where appropriate.

3.2. Session variations and improvement

The variations of physiological metrics: ΔM (pupil diameter), ΔE (EI), and ΔG (gaze entropy) were hypothesized to differ between the two improvement categories. ANOVA models were used to compare the means between the two categories: improvement and no improvement. There were significant differences between the improvement categories for ΔE ($F_{1,120} = 10.02$, $p = .002$) and ΔG ($F_{1,120} = 21.75$, $p < .001$), but not for ΔM. The mean and standard error of session variations under the two categories are shown in Figure 4. In the no improvement group, the means for ΔE and ΔG were both > 0, i.e., the measurement increased since the last session. In the improvement group, ΔE and ΔG were negative, i.e., the measurement decreased since the last session.

3.3. Cognitive/Behavioral metrics and performance

Repeated measures correlation tests showed that both ΔE ($r_{rm} = -.33$, $p < .001$) and ΔG (r_{rm} – .37, $p < .001$) correlated with ΔP; the effect size was medium for both metrics (Bakdash & Marusich, 2017). The negative correlation indicated that large increase in performance was accompanied by larger decrease in EI and Gaze entropy (Figure 5). ΔM was not correlated with ΔP.

3.4. Classification of improvement

In addition to measurements above, time interval between sessions showed significant correlation with ΔP. Yet none of the demographic variables showed a significant relationship with ΔP (based on correlation and ANOVA test). Therefore, four features were used as input: the three metric session variations (ΔM, ΔE and ΔG) and the time interval between sessions. The total number of 122 observations were partitioned into 3 sets with the size of 41, 41 and 40 for the k-fold cross-validation. Of the three tested classifiers, the Naïve Bayes classifier achieved the best performance with an average accuracy of 72.2% (Table 3).

3.5. Subjective metrics variation

To compare the results for ΔE and ΔM, we examined subjective metrics also for H1. As shown in Figure 6, all three subjective metrics differed between improvement and no-improvement groups: $\Delta NASA$ ($F_{1,120} = 55.18$, $p < .001$), $\Delta NASA^M$ ($F_{1,120} = 11.60$, $p < .001$) and $\Delta NASA^E$ ($F_{1,120} = 36.51$, $p < .001$).

All three subjective metrics negatively correlated with performance change: $\Delta NASA$ ($r_{rm} = -.51$, $p < .001$), $\Delta NASA^M$($r_{rm} = -.27$, $p < .001$), $\Delta NASA^E$ ($r_{rm} = -.43$, $p < .001$). Raw-TLX subscale had a large effect size, effort subscale had a medium effect size while the mental demand subscale had a small effect size.

4. Discussion

Robotic surgery is a surgical technique that is gaining popularity, and simulationbased training is a validated and commonly used approach for surgeons to gain the skills needed to perform this advanced technique. Current simulator training is primarily guided by task performance scores with limited consideration of user cognitive states. Physiological and behavioral measurements can provide additional individualized measures for guiding

simulator training. Focusing on changes between training sessions, two hypotheses were tested to explore the relationship between physiological/behavioral responses on performance changes.

4.1. Cognitive/Behavioral metrics and performance

Three metrics were examined in this experiment. Engagement, as measured by EEGbased index, reflected participants' sustained attention to a task (Chaouachi & Frasson, 2010; Freeman, Mikulka, Prinzel, & Scerbo, 1999; Freeman, Mikulka, Scerbo, & Scott, 2004). Previous studies on engagement and performance have suggested that performance is better when operators pay more attention to the task, as summarized in Table 4. These studies have focused on comparing performance between different participants while our experiment focused on how the same participant changed with repetitive training. Therefore, our findings showed that engagement could decrease as participant became proficient at a specific robotic task with multiple practice sessions, instead of how engagement impacted performance at one point in time. This can be understood from the perspective of skill acquisition theory: It takes a series of sequenced stages for initial representation of knowledge to become highly skilled behavior. One important process is the restructuring of procedural knowledge (knowing how to perform a process), which results in automatized knowledge (process performed correctly and rapidly) (DeKeyser, 2007). The general agreement is that with more automatized knowledge, users will require less attention the task process and will be less error-prone (Dekeyser & Criado, 2012).

Gaze entropy represents the distribution of gaze points over a certain period, with a sparser distribution resulting in higher entropy value. Previous use of the gaze entropy metric was mostly in mental workload studies: the concurrence of high workload and high gaze entropy reflected the situation when existing visual exploration strategies could not meet the task requirement (Di Stasi et al., 2016, 2017; Wu et al., 2019). Visual searching plays an important role in robotic surgery for surgeons need to locate targeted organs/issues and respond to visual feedback. More experienced surgeons could potentially locate targets in a faster manner without looking at irrelevant areas. Our study linked this metric with improvement between sessions and demonstrated that gaze entropy can be used for quantifying the effect of training on surgeons' visual search skills.

Mental workload did not show a significant relationship with performance, yet results for subjective measures indicated that changes in the NASA-TLX were highly correlated with changes in performance, meaning that perceived workload decreased when there was a learning effect. This study used pupil diameter as indicator for mental workload, but findings from the NASA-TLX measure of workload suggest that pupil diameter may not be adequately capturing perceived workload during RAS training. Studies reported that the pupil dilates during mental activities (Ambler et al., 1977; Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Granholm & Steinhauer, 2004), but visual stimuli are also have substantial influence on pupil size (Barbur et al., 1992). The robotic task environment contains various visual stimuli like color changes, light changes and moving objects, which could have been a covariate to workload's effect on pupil dilation. For example, participants who experienced high workload tended to make mistakes when objects were flashing or dropped out of view,

which can lead to pupillary constriction. There are a number of other physiological metrics that have been used as measurements for mental workload (Marinescu et al., 2018), and alternative measurements should be examined in future study to identify other objective sensing approach to automatically monitor workload.

4.2. Implications for RAS training

The significant yet low correlations between engagement index, gaze entropy and performance may potentially support our assumption that these metrics are all relevant to surgical skill between sessions, although each provides unique information that is not explained by performance alone. Factors that influence performance changes can be multifaceted and may not be fully captured by metrics in this study; these metrics may therefore provide additional information that can supplement current training assessment. Literature showed that experienced surgeons outperformed novices in motor skills, cognitive states and visual searching, but practice time on a simulator alone cannot quantify improvements on all aspects. The performance score based on motor skills reflects one dimension of surgical performance, yet novices could need significantly more attention/engagement than experts despite similar motor behaviors (Shetty et al., 2016). High engagement is regarded as an indicator that the task is difficult for operators (Berka et al., 2007; Galin et al., 1978; Rabbi et al., 2012) and engagement measures can potentially inform the perceived training difficulty for RAS. Likewise, visual searching is another important aspect of surgery which may not be reflected in motor skills. Recent studies have suggested that gaze pattern of experts should be used in training of novice surgeons (Chetwood et al., 2012; Merali et al., 2017; Tien et al., 2014). This study therefore provides a multidimensional way of evaluating surgical performance that is not captured by just practice time or performance score.

To translate findings to practice, one approach is incorporating the sensing metrics to trainee feedback. For example, an overall score combining all metrics and performance score can improve the current score as a criterion for training outcomes. While the EEG device still requires additional setup, an eye-tracking system can actually be embedded in simulators to provide real-time feedback (Asan & Yang, 2015). Therefore, if trainees improve only their motor skills but not their attention allocation and visual searching skills, a combined score can better reflect the outcome than the one-dimensional score. The training mentor can make judgment accordingly for whether the trainee had skillful operations and efficient visual searching but also whether the trainee required high cognitive load to accomplish the observed technical performance. Although future research is needed to determine best way to deliver feedback for these different metrics, these metrics can potential complement the daily practice for training of RAS.

The findings can help towards addressing some of the key challenges in current RAS training. For example, the technical difficulties can pose high workload for training and for live surgery, and our results demonstrate the feasibility of measuring cognitive states during training. By monitoring engagement and workload, evaluators can ascertain that trainees can perform tasks well with appropriate mental capacity, which lowers the possibility of errors when they transit into real surgery. The real-time data processing ability of EEG and eye

tracking provides the opportunity for more specific diagnosis of training problems. The additional metrics can also address the limitation of one-dimensional evaluation in current simulators which can improve validity of the RAS simulator. The validity of the virtual reality simulation for RAS training has been well-documented in prior research (Bric et al., 2016), and the engagement and gaze entropy metrics add additional evidence for construct validity. The information from measuring trainees' cognitive states and visual behaviors provides additional insight into understanding the overall performance from the simulator exercises, and therefore, whether trainees have gained the necessary surgical skills.

Furthermore, these metrics help understand how skills learned in the simulation relate to those needed in actual surgery. For example, the performance score is specific to computer-simulated tasks and might not available in other training scenarios and live surgery. We examined the alternative of using cognitive and behavioral measures to replace this score. Through supervised machine learning an average accuracy of 72.5% was achieved in predicting improvement. While there were few similar studies we could compare this result with, previous studies that investigated the possibility of classifying workload with physiological signals reported accuracy above 80% as satisfying (Le, Aoki, Murase, & Ishida, 2018; Wilson & Russell, 2003b, 2003a). Therefore, future studies are needed to improve the diagnostic accuracy for this prediction approach. With better accuracy, this alternative could be helpful for continuously evaluating surgeons when they transition from simulated training to live surgery as both EEG and eye-tracking measurements are feasible in the operation room. Future studies should also compare these metrics with subjective ratings for surgical performance (Goh et al., 2012) to determine the application in real surgery.

4.3. Limitations

There are some limitations in the current study. Due to the curriculum nature, task order was not randomized, which might produce a confounded order effect. To randomize the task order, future research may consider recruiting participants who have already gained basic skills and do not need to follow the skills progression outlined by the task order. In addition, the number of training sessions were limited by trainee's schedule, resulting in learning curves that may not have yet reached the plateaued stage (especially for the more advanced tasks/difficulties). Furthermore, due to constraints in trainees' surgical schedules and the availability of the simulator (which is used for live surgeries), time elapsed between training sessions varied. Due to differences in individual reactions to training and physiological measurements, the small number of participants can potentially impact generalizability of the results. Although significant relationships were still identified with this limitation, future studies with large sample size and more controlled training interval should investigate how these findings relate to different stages of learning and total practice time. Finally, the da Vinci simulator used in this study is commercially available and used worldwide, which supports the validity of tasks and relevance to real training process. However, as a commercial product, it also restricted our freedom to modify task elements, explore the impact of task structure, and fully understanding the task factors behind the performance score. Therefore, more customized simulations and software that enhance experimental control could consolidate and extend the current findings.

5. Conclusion

Robotic surgery has been increasingly adopted as an advanced medical technique, and well-designed training curriculum is needed to ensure the safety and effectiveness of its usage. This study provides a user-sensing perspective to simulated training: by focusing on changes between training sessions, we can identify how learning impacts trainee's cognitive states and behaviors. Findings demonstrate that effective robotic training has resulted in fewer mental resources allocated to the task and improved visual searching strategy. It provides insight for how surgeons' engagement, gaze entropy and mental workload changes as they gain skills in robotic surgery. This demonstration allows the possibility of using real-time measurement to monitor the training process and providing assessments other than task performance score.

6.: Appendix

An alternative way to categorize improvement is through the 1/3 and 3/3 quantile. Results under this approach are shown below. Results for H2 are not reported since improvement labels were not involved in the correlation analyses.

6.1. Performance variation

Figure A.1 shows the result of categorization. 73 observations with ΔP no larger than 0 were categorized as non-improvement and 73 observations with ΔP no less than 10 were categorized as improvement.

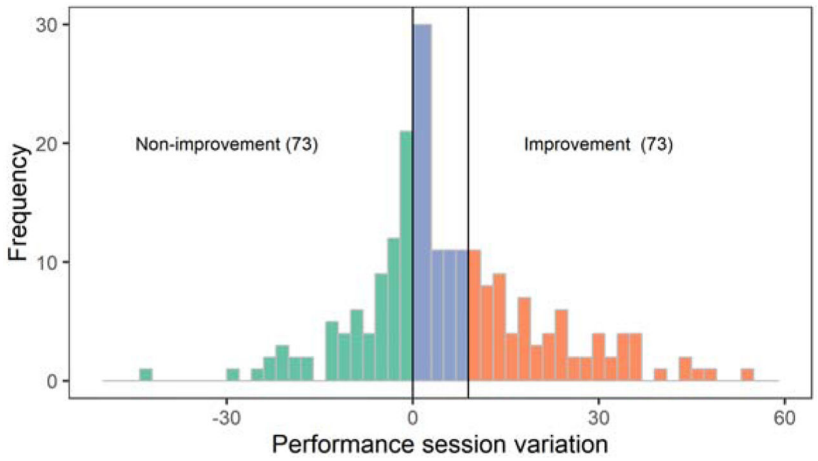

Figure A.1.
Histogram of performance change

6.2. Session variations and improvement

F tests indicated there were significant difference between conditions for ΔE ($F_{1,144} = 15.68$, $p < .001$) and ΔG ($F_{1,144} = 19.11$, $p < .001$).

6.3. Classification of improvement

The 146 observations were partitioned into 3 sets with the size of 48, 48 and 50 for the k-fold cross-validation. The average accuracy was 66.9% and the average F1 score 0.63.

6.4. Subjective metrics variation

For H1, ANOVA tests indicated that sessions variations for subjective metrics were different in the two categories: $\Delta NASA$ ($F_{1,144} = 54.88$, $p < .001$), $\Delta NASA^M$ ($F_{1,144} = 15.67$, $p < .001$) and $\Delta NASA^E$ ($F_{1,144} = 66.45$, $p < .001$).

Reference

Al Shalabi L, Shaaban Z, & Kasasbeh B (2006). Data mining: A preprocessing engine. Journal of Computer Science, 2(9), 735–739.

Alzahrani T, Haddad R, Alkhayal A, Delisle J, Drudi L, Gotlieb W, Fraser S, Bergman S, Bladou F, Andonian S, & Anidjar M (2013). Validation of the da Vinci Surgical Skill Simulator across three surgical disciplines: A pilot study. Canadian Urological Association Journal, 7(7–8), E520–E529. 10.5489/cuaj.419 [PubMed: 23914275]

Ambler BA, Fisicaro SA, & Proctor RW (1977). Information reduction, internal transformations, and task difficulty. Bulletin of the Psychonomic Society, 10(6), 463–466. 10.3758/BF03337698

Asan O, & Yang Y (2015). Using Eye Trackers for Usability Evaluation of Health Information Technology: A Systematic Literature Review. JMIR Human Factors, 2(1), e5 10.2196/humanfactors.4062 [PubMed: 27026079]

Bakdash JZ, & Marusich LR (2017). Repeated Measures Correlation. Frontiers in Psychology, 8, 456 10.3389/fpsyg.2017.00456 [PubMed: 28439244]

Barbur JL, Harlow AJ, & Sahraie A (1992). Pupillary responses to stimulus structure, colour and movement. Ophthalmic and Physiological Optics, 12(2), 137–141. 10.1111/j.1475-1313.1992.tb00276.x [PubMed: 1408159]

Batista GEAPA, Prati RC, & Monard MC (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, 6(1), 20–29. 10.1145/1007730.1007735

Beatty J (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological Bulletin, 91(2), 276. [PubMed: 7071262]

Beatty J, & Kahneman D (1966). Pupillary changes in two memory tasks. Psychonomic Science, 5(10), 371–372.

Beatty J, & Lucero-Wagoner B (2000). The pupillary system. Handbook of Psychophysiology, 2(142–162).

Benitez DS, Toscano S, & Silva A (2016). On the use of the Emotiv EPOC neuroheadset as a low cost alternative for EEG signal acquisition. 2016 IEEE Colombian Conference on Communications and Computing (COLCOM), 1–6. 10.1109/ColComCon.2016.7516380

Benjamini Y, & Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300.

Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT, Olmstead RE, Tremoulet PD, & Craven PL (2007). EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. 78(5), 14.

Bric JD, Lumbard DC, Frelich MJ, & Gould JC (2016). Current state of virtual reality simulation in robotic surgery training: A review. Surgical Endoscopy, 30(6), 2169–2178. 10.1007/s00464-015-4517-y [PubMed: 26304107]

Brinkman WM, Luursema J-M, Kengen B, Schout BMA, Witjes JA, & Bekkers RL (2013). Da Vinci Skills Simulator for Assessing Learning Curve and Criterionbased Training of Robotic Basic Skills. Urology, 81(3), 562–566. 10.1016/j.urology.2012.10.020 [PubMed: 23295136]

Cain B (2007). A review of the mental workload literature. Defence Research And Development Toronto (Canada).

Carswell CM, Clarke D, & Seales WB (2005). Assessing mental workload during laparoscopic surgery. Surgical Innovation, 12(1), 80–90. [PubMed: 15846451]

Chaouachi M, & Frasson C (2010). Exploring the Relationship between Learner EEG Mental Engagement and Affect In Aleven V, Kay J, & Mostow J (Eds.), Intelligent Tutoring Systems (Vol. 6095, pp. 291–293). Springer Berlin Heidelberg 10.1007/978-3-642-13437-1_48

Charles RL, & Nixon J (2019). Measuring mental workload using physiological measures: A systematic review. Applied Ergonomics, 74, 221–232. 10.1016/j.apergo.2018.08.028 [PubMed: 30487103]

Chawla NV, Japkowicz N, & Kotcz A (2004). Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6(1), 1–6. 10.1145/1007730.1007733

Chetwood AS, Kwok K-W, Sun L-W, Mylonas GP, Clark J, Darzi A, & Yang GZ (2012). Collaborative eye tracking: A potential training tool in laparoscopic surgery. Surgical Endoscopy, 26(7), 2003–2009. [PubMed: 22258302]

Coelli S, Sclocco R, Barbieri R, Reni G, Zucca C, & Bianchi AM (2015). EEGbased index for engagement level monitoring during sustained attention. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 1512–1515. 10.1109/EMBC.2015.7318658

Da Vinci S System User Manual (User Manual No. 550516–06). (2014). Intuitive Surgical, Inc.

DeKeyser R (2007). Skill acquisition theory. Theories in Second Language Acquisition: An Introduction, 97113.

Dekeyser R, & Criado R (2012). Automatization, Skill Acquisition, and Practice in Second Language Acquisition. In The Encyclopedia of Applied Linguistics. American Cancer Society. 10.1002/9781405198431.wbeal0067

Delorme A, & Makeig S (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of Neuroscience Methods, 134(1), 9–21. 10.1016/j.jneumeth.2003.10.009 [PubMed: 15102499]

Di Nocera F, Camilli M, & Terenzi M (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. Journal of Cognitive Engineering and Decision Making, 1(3), 271–285.

Di Stasi LL, Diaz-Piedra C, Rieiro H, Carrión JMS, Berrido MM, Olivares G, & Catena A (2016). Gaze entropy reflects surgical task load. Surgical Endoscopy, 30(11), 5034–5043. 10.1007/s00464-016-4851-8 [PubMed: 26983440]

Di Stasi LL, Díaz-Piedra C, Ruiz-Rabelo JF, Rieiro H, Sanchez Carrion JM, & Catena A (2017). Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based indices. Applied Ergonomics, 65, 168–74. 10.1016/j.apergo.2017.06.008 [PubMed: 28802436]

Diana M, & Marescaux J (2015). Robotic surgery, British Journal of Surgery., 102(2), e15–e28. 10.1002/bjs.9711 [PubMed: 25627128]

Fawcett T (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. 10.1016/j.patrec.2005.10.010

Finnegan KT, Meraney AM, Staff I, & Shichman SJ (2012). da Vinci Skills Simulator Construct Validation Study: Correlation of Prior Robotic Experience With Overall Score and Time Score Simulator Performance. Urology, 80(2), 330–336. 10.1016/j.urology.2012.02.059 [PubMed: 22704177]

Foxe JJ, & Snyder AC (2011). The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. Frontiers in Psychology, 2 10.3389/fpsyg.2011.00154

Freeman FG, Mikulka PJ, Prinzel LJ, & Scerbo MW (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. Biological Psychology, 50(1), 61–76. 10.1016/S0301-0511(99)00002-2 [PubMed: 10378439]

Freeman FG, Mikulka PJ, Scerbo MW, & Scott L (2004). An evaluation of an adaptive automation system using a cognitive vigilance task. Biological Psychology, 67(3), 283–297. 10.1016/j.biopsycho.2004.01.002 [PubMed: 15294387]

Galin D, Johnstone J, & Herron J (1978). Effects of task difficulty on EEG measures of cerebral engagement. Neuropsychologia, 16(4), 461–472. 10.1016/0028-3932(78)90069-6 [PubMed: 692858]

Gardner AK, Jabbour IJ, Williams BH, & Huerta S (2016). Different Goals, Different Pathways: The Role of Metacognition and Task Engagement in Surgical Skill Acquisition. Journal of Surgical Education, 73(1), 61–65. 10.1016/j.jsurg.2015.08.007 [PubMed: 26395402]

Giulianotti PC, Coratti A, Sbrana F, Addeo P, Bianco FM, Buchs NC, Annechiarico M, & Benedetti E (2011). Robotic liver surgery: Results for 70 resections. Surgery, 149(1), 29–39. 10.1016/j.surg.2010.04.002 [PubMed: 20570305]

Goh AC, Goldfarb DW, Sander JC, Miles BJ, & Dunkin BJ (2012). Global Evaluative Assessment of Robotic Skills: Validation of a Clinical Assessment Tool to Measure Robotic Surgical Skills. The Journal of Urology, 187(1), 247–252. 10.1016/j.juro.2011.09.032 [PubMed: 22099993]

Gola M, Magnuski M, Szumska I, & Wróbel A (2013). EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects. International Journal of Psychophysiology, 89(3), 334–341. 10.1016/j.ijpsycho.2013.05.007 [PubMed: 23688673]

Granholm E, & Steinhauer SR (2004). Pupillometric measures of cognitive and emotional processes. International Journal of Psychophysiology, 52(1), 1–6. 10.1016/j.ijpsycho.2003.12.001 [PubMed: 15003368]

Guru KA, Esfahani ET, Raza SJ, Bhat R, Wang K, Hammond Y, Wilding G, Peabody JO, & Chowriappa AJ (2015). Cognitive skills assessment during robot-assisted surgery: Separating the wheat from the chaff. BJU International, 115(1), 166–174. 10.1111/bju.12657 [PubMed: 24467726]

Guru KA, Shafiei SB, Khan A, Hussein AA, Sharif M, & Esfahani ET (2015). Understanding Cognitive Performance During Robot-Assisted Surgery. Urology, 86(4), 751–757. 10.1016/j.urology.2015.07.028 [PubMed: 26255037]

Hart SG (2006). NASA-task load index (NASA-TLX); 20 years later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50, 904–908.

Hart SG, & Staveland LE (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research In Advances in psychology (Vol. 52, pp. 139–183). Elsevier.

Hastie T, Friedman J, & Tibshirani R (2001). Model Assessment and Selection In The Elements of Statistical Learning (pp. 193–224). Springer, New York, NY 10.1007/978-0-387-21606-5_7

Hermens DF, Soei EXC, Clarke SD, Kohn MR, Gordon E, & Williams LM (2005). Resting EEG theta activity predicts cognitive performance in attention-deficit hyperactivity disorder. Pediatric Neurology, 32(4), 248–256. 10.1016/j.pediatrneurol.2004.11.009 [PubMed: 15797181]

Horrey WJ, Lesch MF, Garabet A, Simmons L, & Maikala R (2017). Distraction and task engagement: How interesting and boring information impact driving performance and subjective and physiological responses. Applied Ergonomics, 58, 342–348. 10.1016/j.apergo.2016.07.011 [PubMed: 27633231]

Hu JSL, Lu J, Tan WB, & Lomanto D (2016). Training improves laparoscopic tasks performance and decreases operator workload. Surgical Endoscopy, 30(5), 1742–1746. 10.1007/s00464-015-4410-8 [PubMed: 26173550]

James G, Witten D, Hastie T, & Tibshirani R (2013). An introduction to statistical learning (Vol. 112). Springer.

Jayalakshmi T, & Santhakumaran A (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, 3(1), 1793–8201.

Jraidi I, Khedher AB, Chaouachi M, & Frasson C (2019). Assessing Students' Clinical Reasoning Using Gaze and EEG Features In Coy A, Hayashi Y, & Chang M (Eds.), Intelligent Tutoring Systems (pp. 47–56). Springer International Publishing 10.1007/978-3-030-22244-4_7

Kamiúki J, Brzezicka A, Gola M, & Wróbel A (2012). Beta band oscillations engagement in human alertness process. International Journal of Psychophysiology, 55(1), 125–128. 10.1016/j.ijpsycho.2011.11.006

Kenney PA, Wszolek MF, Gould JJ, Libertino JA, & Moinzadeh A (2009). Face, Content, and Construct Validity of dV-Trainer, a Novel Virtual Reality Simulator for Robotic Surgery. Urology, 73(6), 1288–1292. 10.1016/j.urology.2008.12.044 [PubMed: 19362352]

Khedher AB, Jraidi I, & Frasson C (2019). Tracking Students' Mental Engagement Using EEG Signals during an Interaction with a Virtual Learning Environment. Journal of Intelligent Learning Systems and Applications, 11(1), 1–14. 10.4236/jilsa.2019.111001

Lai F, & Entin E (2005). Robotic Surgery and the Operating Room Team. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 49(11), 1070–1073. 10.1177/154193120504901115

Lanfranco AR, Castellanos AE, Desai JP, & Meyers WC (2004). Robotic Surgery. Annals of Surgery, 239(1), 14–21. 10.1097/01.sla.0000103020.19595.7d [PubMed: 14685095]

Le AS, Aoki H, Murase F, & Ishida K (2018). A Novel Method for Classifying Driver Mental Workload Under Naturalistic Conditions With Information From Near-Infrared Spectroscopy. Frontiers in Human Neuroscience, 12 10.3389/fnhum.2018.00431

Lelis-Torres N, Ugrinowitsch H, Apolinário-Souza T, Benda RN, & Lage GM (2017). Task engagement and mental workload involved in variation and repetition of a motor skill. Scientific Reports, 7(1), 1–10. 10.1038/s41598-017-15343-3 [PubMed: 28127051]

Lerner MA, Ayalew M, Peine WJ, & Sundaram CP (2010). Does Training on a Virtual Reality Robotic Simulator Improve Performance on the da Vinci® Surgical System? Journal of Endourology, 24(3), 467–472. 10.1089/end.2009.0190 [PubMed: 20334558]

Mack MJ (2001). Minimally Invasive and Robotic Surgery. The Journal of the American Medical Association, 285(5), 568–572. 10.1001/jama.285.5.568 [PubMed: 11176860]

Marinescu AC, Sharples S, Ritchie AC, Sánchez López T, McDowell M, & Morvan HP (2018). Physiological Parameter Response to Variation of Mental Workload. Human Factors, 60(1), 31–56. 10.1177/0018720817733101 [PubMed: 28965433]

Marshall SP (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity.

Matthews G, Warm JS, & Smith AP (2017). Task Engagement and Attentional Resources: Multivariate Models for Individual Differences and Stress Factors in Vigilance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 59(1), 44–61. 10.1177/0018720816673782

Merali N, Veeramootoo D, & Singh S (2017). Eye-Tracking Technology in Surgical Training. Journal of Investigative Surgery, 1–7.

Mognon A, Jovicich J, Bruzzone L, & Buiatti M (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology, 48(2), 229–240. 10.1111/j.1469-8986.2010.01061.x [PubMed: 20636297]

Morris B (2005). Robotic Surgery: Applications, Limitations, and Impact on Surgical Education. Medscape General Medicine, 7(3), 72.

Mosaly PR, Mazur LM, & Marks LB (2017). Quantification of baseline pupillary response and task-evoked pupillary response during constant and incremental task load. Ergonomics, 60(10), 1369–1375. 10.1080/00140139.2017.1288930 [PubMed: 28140793]

Mullen T (2012). CleanLine EEGLAB plugin. San Diego, CA: Neuroimaging Informatics Toolsand Resources Clearinghouse (NITRC).

Narazaki K, Oleynikov D, & Stergiou N (2006). Robotic surgery training and performance. Surgical Endoscopy And Other Interventional Techniques, 20(1), 96–103. 10.1007/s00464-005-3011-3 [PubMed: 16374675]

Palinko O, Kun AL, Shyrokov A, & Heeman P (2010). Estimating cognitive load using remote eye tracking in a driving simulator. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, 141–144.

Patten CJD, Kircher A, Östlund J, Nilsson L, & Svenson O (2006). Driver experience and cognitive workload in different traffic environments. Accident Analysis & Prevention, 38(5), 887–894. 10.1016/j.aap.2006.02.014 [PubMed: 16620740]

Perrenot C, Perez M, Tran N, Jehl J-P, Felblinger J, Bresler L, & Hubert J (2012). The virtual reality simulator dV-Trainer® is a valid assessment tool for robotic surgical skills. Surgical Endoscopy, 26(9), 2587–2593. 10.1007/s00464-012-2237-0 [PubMed: 22476836]

Pomplun M, & Sunkara S (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. Proceedings of the International Conference on HCI, 2003.

Pope AT, Bogart EH, & Bartolome DS (1995). Biocybernetic system evaluates indices of operator engagement in automated task. Biological Psychology, 40(1), 187–195. 10.1016/0301-0511(95)05116-3 [PubMed: 7647180]

Posner MI (1988). Structures and function of selective attention In Clinical neuropsychology and brain function: Research, measurement, and practice (pp. 173–202). American Psychological Association 10.1037/10063-005

Posner MI, & Petersen SE (1990). The Attention System of the Human Brain. Annual Review of Neuroscience, 13(1), 25–42. 10.1146/annurev.ne.13.030190.000325

Rabbi AF, Zony A, de Leon P, & Fazel-Rezai R (2012). Mental workload and task engagement evaluation based on changes in electroencephalogram. Biomedical Engineering Letters, 2(3), 139–146. 10.1007/s13534-012-0065-8

Rassweiler J, Binder J, & Frede T (2001). Robotic and telesurgery: Will they change our future? Current Opinion in Urology, 11(3), 309–320. [PubMed: 11371786]

Ruiz-Rabelo JF, Navarro-Rodriguez E, Di-Stasi LL, Diaz-Jimenez N, Cabrera-Bermon J, Diaz-Iglesias C, Gomez-Alvarez M, & Briceño-Delgado J (2015). Validation of the NASA-TLX Score in Ongoing Assessment of Mental Workload During a Laparoscopic Learning Curve in Bariatric Surgery. Obesity Surgery, 25(12), 2451–2456. 10.1007/s11695-015-1922-1 [PubMed: 26459432]

Sackier JM, & Wang Y (1994). Robotically assisted laparoscopic surgery. Surgical Endoscopy, 8(1), 63–66. 10.1007/BF02909496 [PubMed: 8153867]

Schiff L, Tsafrir Z, Aoun J, Taylor A, Theoharis E, & Eisenstein D (2016). Quality of Communication in Robotic Surgery and Surgical Outcomes. JSLS : Journal of the Society of Laparoendoscopic Surgeons, 20(3). 10.4293/JSLS.2016.00026

Sharma A, & Singh M (2015). Assessing alpha activity in attention and relaxed state: An EEG analysis. 2015 1st International Conference on Next Generation Computing Technologies (NGCT), 508–513. 10.1109/NGCT.2015.7375171

Shetty K, Leff DR, Orihuela-Espina F, Yang G-Z, & Darzi A (2016). Persistent Prefrontal Engagement Despite Improvements in Laparoscopic Technical Skill. JAMA Surgery, 151(7), 682–684. 10.1001/jamasurg.2016.0050 [PubMed: 27028901]

Singleton RC (1979). Mixed Radix Fast Fourier Transforms, in Programs for Digital Signal Processing. IEEE Digital Signal Processing Committee, Eds., IEEE Press.

So WKY, Wong SWH, Mak JN, & Chan RHM (2017). An evaluation of mental workload with frontal EEG. PLOS ONE, 12(4), e0174949 10.1371/journal.pone.0174949 [PubMed: 28414729]

Solovey ET, Zec M, Garcia Perez EA, Reimer B, & Mehler B (2014). Classifying driver workload using physiological and driving performance data: Two field studies. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 4057–4066.

Stytsenko K, Jablonskis E, & Prahm C (2011, 6 8). Evaluation of consumer EEG device Emotiv EPOC. MEi:CogSci Conference 2011, Ljubljana MEi:CogSci Conference 2011, Ljubljana. https://www.univie.ac.at/meicogsci/php/ocs/index.php/meicog/meicog2011/paper/view/210

Tien T, Pucher PH, Sodergren MH, Sriskandarajah K, Yang G-Z, & Darzi A (2014). Eye tracking for skills assessment and training: A systematic review. Journal of Surgical Research, 191(1), 169–178. 10.1016/j.jss.2014.04.032 [PubMed: 24881471]

Venables L, & Fairclough SH (2009). The influence of performance feedback on goalsetting and mental effort regulation. Motivation and Emotion, 55(1), 63–74. 10.1007/s11031-008-9116-y

Wascher E, Rasch B, Sänger J, Hoffmann S, Schneider D, Rinkenauer G, Heuer H, & Gutberlet I (2014). Frontal theta activity reflects distinct aspects of mental fatigue. Biological Psychology, 96, 57–65. 10.1016/j.biopsycho.2013.11.010 [PubMed: 24309160]

Wickens CD (2008). Multiple Resources and Mental Workload. Human Factors, 50(3), 449–455. 10.1518/001872008X288394 [PubMed: 18689052]

Wilson GF, & Russell CA (2003a). Operator Functional State Classification Using Multiple Psychophysiological Features in an Air Traffic Control Task. Human Factors, 45(3), 381–389. 10.1518/hfes.45.3.381.27252 [PubMed: 14702990]

Wilson GF, & Russell CA (2003b). Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. Human Factors, 45(4), 635–644. 10.1518/hfes.45.4.635.27088 [PubMed: 15055460]

Wilson MR, Poolton JM, Malhotra N, Ngo K, Bright E, & Masters RSW (2011). Development and Validation of a Surgical Workload Measure: The Surgery Task Load Index (SURG-TLX). World Journal of Surgery, 35(9), 1961 10.1007/s00268-011-1141-4 [PubMed: 21597890]

Wu C, Cha J, Sulek J, Zhou T, Sundaram CP, Wachs J, & Yu D (2019). Eye-Tracking Metrics Predict Perceived Workload in Robotic Surgical Skills Training. Human Factors, 0018720819874544. 10.1177/0018720819874544

Yu D, Lowndes B, Thiels C, Bingener J, Abdelrahman A, Lyons R, & Hallbeck S (2016). Quantifying Intraoperative Workloads Across the Surgical Team Roles: Room for Better Balance? World journal of surgery, 40(7), 1565–1574. doi:10.1007/s00268-016-3449-6 [PubMed: 26952115]

Zhang J-Y, Liu S-L, Feng Q-M, Gao J-Q, & Zhang Q (2017). Correlative Evaluation of Mental and Physical Workload of Laparoscopic Surgeons Based on Surface Electromyography and Eye-tracking Signals. Scientific Reports, 7(1), 11095 10.1038/s41598-017-11584-4 [PubMed: 28894216]

Zheng B, Jiang X, & Atkins MS (2015). Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses. Surgical Innovation, 22(6), 629–635. 10.1177/1553350615573582 [PubMed: 25759398]

Highlights

- Participants had varied performance changes between training sessions
- Changes in performance were negatively correlated with changes in engagement index and gaze entropy
- Cognitive and behavioral metrics predict training outcomes with 72.5% accuracy
- Sensor-based metrics can complement current feedback tools for predicting training performance

Figure 1.
Participant performing tasks through the robotic console

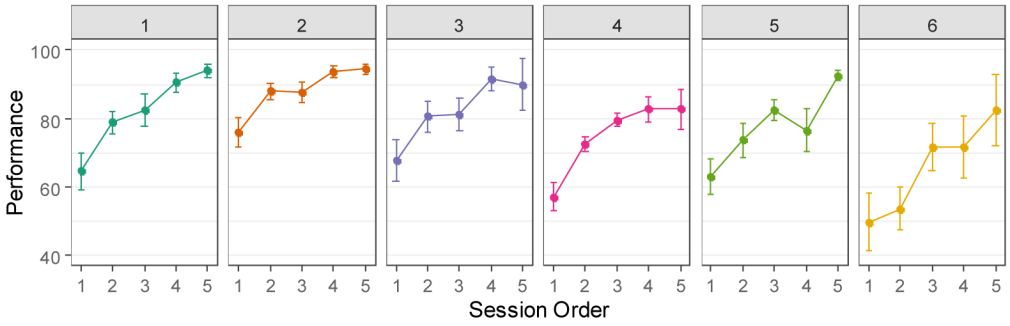

Figure 2.
Average performance trend for six tasks (with standard error)

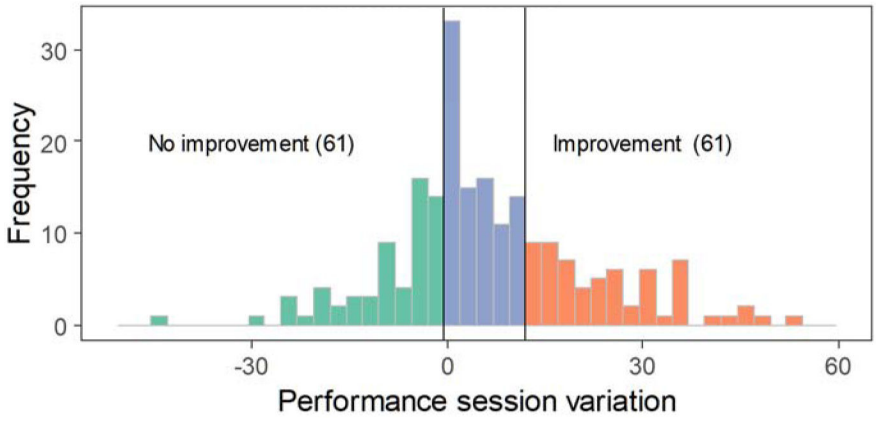

Figure 3.
Histogram of performance change (n=212 total observations)

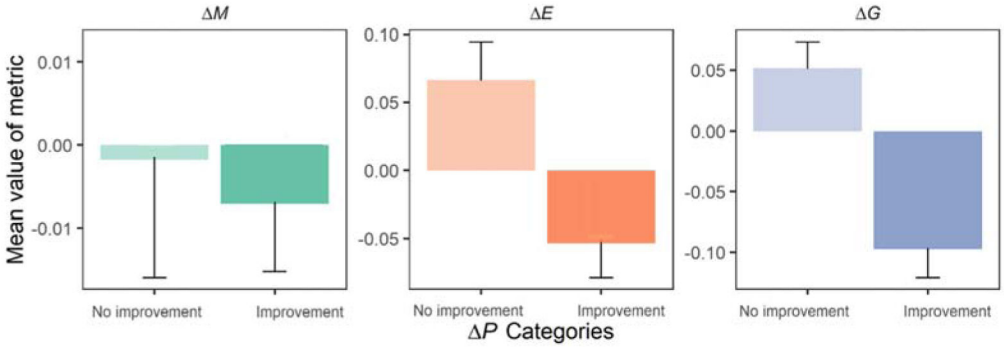

Figure 4.

Bar plot of the mean value of three metrics: Δ*M* (pupil diameter), Δ*E* (EI) and Δ*G* (gaze entropy), x-axis divided by two categories of improvement. Standard error added as error bar.

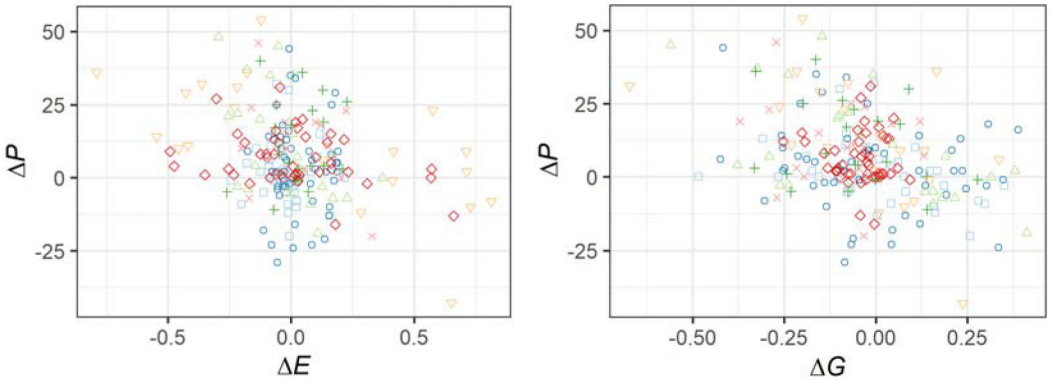

Figure 5.
Correlation between session variations in performance (*ΔP*) and cognitive/behavioral metrics (*ΔE*, *ΔG*) (colored and shaped by participants)

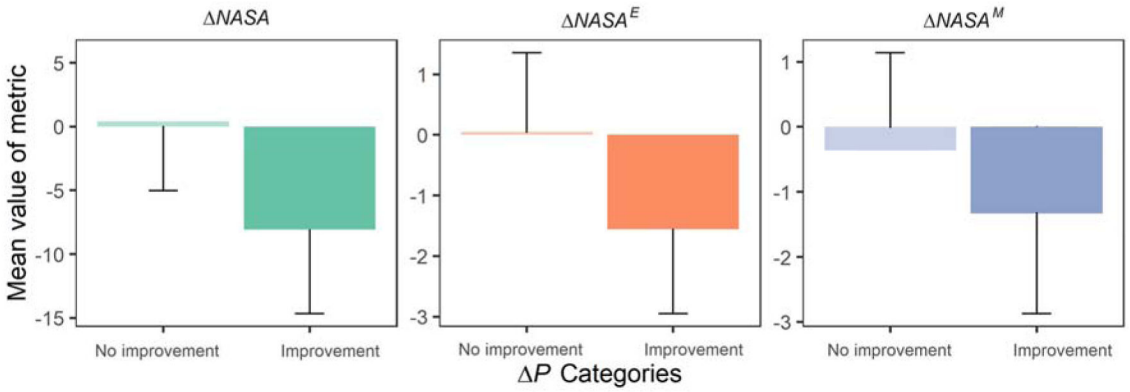

Figure 6.

Bar plot of the mean value of three metrics: $\Delta NASA$ (Raw TLX), $\Delta NASA^{E}$ (Effort) and $\Delta NASA^{M}$ (Mental Demand)

Table 1

Simulated training tasks

	Task description
1	Camera Targeting (CT): Focus the camera on different blue spheres spread across a broad pelvic cavity. Two levels of difficulty.
2	Peg Board (PB): Grasp rings on a vertical stand with the left hand and then passing them to the right hand before placing them on a peg. Two levels of difficulty.
3	Ring and Rail (RR): Move a ring along a twisted metal rod without applying excessive force to either the ring or the rail. Two levels of difficulty.
4	Suture Sponge (SS): Drive needle through random targets on a deformable structure. Three levels of difficulty.
5	Dots and Needles (DN): Insert a needle through several pairs of targets that have various spatial positions. Two levels of difficulty.
6	Tubes (T): Drive needle through fixed targets on a cylindrical deformable structure. One level of difficulty.

Table 2

List and definition of metrics

Metrics	Symbol	Calculation
Performance change	ΔP	$\Delta P_{i,j,k} = P_{i,j,k} - P_{i,j,k\text{-}1}$
Pupil diameter (mental workload) change	Δ_{M}	$\Delta M_{i,j,k} = M_{i,j,k} - M_{i,j,k\text{-}1}$
EI (engagement) change	ΔE	$\Delta E_{i,j,k} = E_{i,j,k} - E_{i,j,k\text{-}1}$
Gaze entropy (gaze pattern) change	ΔG	$\Delta S_{i,j,k} = S_{i,j,k} - S_{i,j,k\text{-}1}$
Raw TLX change	$\Delta NASA$	$\Delta NASA_{i,j,k} = NASA_{i,j,k} - NASA_{i,j,k\text{-}1}$
NASA-TLX mental demand subscale change	$\Delta NASA^M$	$\Delta NASA^{M}_{i,j,k} = NASA^{M}_{i,j,k} - NASA^{M}_{i,j,k-1}$
NASA-TLX effort subscale change	$\Delta NASA^E$	$\Delta NASA^{E}_{i,j,k} = NASA^{E}_{i,j,k} - NASA^{E}_{i,j,k-1}$

Table 3

Confusion matrix results for classifying improvement and no-improvement groups using sensing metrics and time

		Actual class		
		No improvement	**Improvement**	
Pridited class	No improvement	32.1%±9.4%	9.8%±2.3%	75.7%±8.1%
		True positive	False positive	Precision
	Improvement	17.9%±9.8%	40.2%±2.0%	70.3%±13.3%
		False negative	True negative	NPV
		80.4%±4.4%	80.4%±4.4%	72.2%±11.3%
		Sensitivity	Specificity	Accuracy

Table 4

Relationship of engagement and performance reported in previous studies

Study	Engagement	Task	Relationship
Khedher (2019)	*Beta/(Alpha+Theta)*	Solving problems in medical cases	Higher engagement in the successful group
Jraidi (2019)	*Beta/(Alpha+Theta)*	Solving problems in medical cases	Disengaged learners had higher number of attempts
Coelli (2015)	*Beta/Alpha*	Signal detection test	Engagement correlated with reaction time [−0.31,−0.26]
Chaouachi (2010)	*Beta/(Alpha+Theta)*	General knowledge questions and spell checking	Learners with higher engagement performed better
Berka (2007)	*Discriminant function analysis of EEG*	Forward/backward digit-span, gridrecall, and mentaladdition tests	Decreased engagement was associated with increased reaction time