

**EYE TRACKING AND ELECTROENCEPHALOGRAPH (EEG)
MEASURES FOR WORKLOAD AND PERFORMANCE IN ROBOTIC
SURGERY TRAINING**

by
Chuhao Wu

A Thesis

*Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of*

Master of Science in Industrial Engineering



School of Industrial Engineering

West Lafayette, Indiana

August 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Denny Yu, Chair

School of Industrial Engineering

Dr. Mark R. Lehto

School of Industrial Engineering

Dr. Robert N. Proctor

School of Industrial Engineering

Dr. Chandru P. Sundaram

India University

Approved by:

Dr. Steven Landry

Head of the Graduate Program

*To people who helped, help, and will help me,
thank you.*

ACKNOWLEDGMENTS

Content of this thesis will be used for journal and conference publications.

I would like to thank all my committee members: Dr. Lehto, Dr. Proctor, and Dr. Sundaram for their patience, support, and valuable feedback. I express gratitude for my advisor, Dr. Denny Yu, who have led me through this long journey towards my degree. His constant guidance and push helped me to stay on track and accomplish this work.

I would like to thank all my friends, colleagues and labmates, for their great support for my research, study, and life at Purdue. I also thank MEERCat Purdue and Dr. Berger for funding my graduate study.

Huge thanks to my family for their unconditional love and support. Thanks to Fish Leong, my favorite singer, for making the world a better place. Finally, thanks to all adversities, for what doesn't kill me makes me stronger.

TABLE OF CONTENTS

| | |
|---|----|
| LIST OF TABLES | 7 |
| LIST OF FIGURES | 8 |
| NOMENCLATURE | 9 |
| ABSTRACT..... | 10 |
| 1. INTRODUCTION | 12 |
| 1.1 Robotic surgery | 12 |
| 1.2 Human factors approach to surgery | 14 |
| 1.3 Structure of the Document | 16 |
| 2. LITERATURE REVIEW | 17 |
| 2.1 Robotic Surgery Training | 17 |
| 2.2 Objective measures in HFE | 19 |
| 2.3 Classification analysis..... | 22 |
| 3. RESEARCH QUESTION I: WORKLOAD MEASURES | 24 |
| 3.1 Research Framework | 24 |
| 3.2 Methodology | 24 |
| 3.2.1 Robotic system and tasks | 24 |
| 3.2.2 Measurement Metrics | 26 |
| 3.2.2.1 Performance Metric | 26 |
| 3.2.2.2 Subjective Metric..... | 27 |
| 3.2.2.3 Eye Tracking Metrics | 27 |
| 3.2.3 Study procedure | 28 |
| 3.2.3.1 Participants | 28 |
| 3.2.3.2 Procedure | 29 |
| 3.2.4 Data Analysis..... | 30 |
| 3.2.4.1 Data scaling | 30 |
| 3.2.4.2 Workload labeling | 30 |
| 3.2.4.3 Statistical Analysis | 31 |
| 3.3 Results..... | 32 |

| | | |
|---------|--|----|
| 3.3.1 | Descriptive data | 32 |
| 3.3.2 | Eye Tracking Metrics and Task Difficulty | 33 |
| 3.3.3 | Eye tracking Metrics and NASA-TLX | 35 |
| 3.3.4 | Classification of workload | 37 |
| 3.4 | Discussion | 39 |
| 4. | RESEARCH QUESTION II: PERFORMANCE MEASURES | 43 |
| 4.1 | Research Framework | 43 |
| 4.2 | Methodology | 43 |
| 4.2.1 | Measurement Metrics | 43 |
| 4.2.2 | Data Analysis | 45 |
| 4.2.2.1 | Session variations | 45 |
| 4.2.2.2 | Improvement Labeling | 45 |
| 4.2.2.3 | Statistical Analysis | 46 |
| 4.3 | Results | 46 |
| 4.3.1 | Descriptive data | 46 |
| 4.3.2 | Objective Metrics and Improvement | 48 |
| 4.3.3 | Objective Metrics and Performance Change | 49 |
| 4.3.4 | Classification of improvement | 49 |
| 4.3.5 | Subjective Metrics | 51 |
| 4.4 | Discussion | 52 |
| 5. | CONCLUSION AND FUTURE RESEARCH | 56 |
| | APPENDIX A. RAS TASK ANALYSIS | 58 |
| | APPENDIX B. NASA-TLX SURVEY | 64 |
| | APPENDIX C. NASA-TLX HISTOGRAMS | 65 |
| | APPENDIX D. TABLE OF CORRELATIONS | 66 |
| | REFERENCES | 67 |

LIST OF TABLES

| | |
|---|----|
| Table 3.1 Training tasks screenshots | 25 |
| Table 3.2 Mean value of all metrics across tasks and levels | 33 |
| Table 3.3 Mixed models summary for effects of task level on eye tracking metrics | 34 |
| Table 3.4 Repeated correlation between subjective metric and eye tracking metrics | 36 |
| Table 3.5 Workload classification confusion matrix for testing dataset..... | 38 |
| Table 4.1 List of session variations..... | 45 |
| Table 4.2 Mean value of all metrics across tasks and attempts | 47 |
| Table 4.3 Objective metrics variation and practice outcome..... | 48 |
| Table 4.4 Improvement classification confusion matrix for testing dataset | 50 |
| Table 4.5 Subjective metrics and improvement..... | 51 |
| Appendix | |
| Table A.1 Simulated robotic surgical tasks analysis | 58 |
| Table A.2 Repeated correlation between eye metrics and NASA-TLX subscales..... | 66 |

LIST OF FIGURES

| | |
|---|----|
| Fig. 1.1 Example of RAS system components | 13 |
| Fig. 1.2 Example of RAS surgeon console interfaces | 14 |
| Fig. 3.1 Screenshot of task performance..... | 26 |
| Fig. 3.2 Participants wearing the eye tracker (left) and performing tasks (right)..... | 30 |
| Fig. 3.3 Mean value of task performance and NASA-TLX across tasks and levels..... | 33 |
| Fig. 3.4 Mean value of eye tracking metrics across tasks and levels..... | 35 |
| Fig. 3.5 Distribution of eye tracking metrics over workload..... | 36 |
| Fig. 3.6 Mean value of eye tracking metrics in high/low workload | 37 |
| Fig. 4.1 Histogram of performance change | 47 |
| Fig. 4.2 Mean values of objective variations in two conditions by task..... | 48 |
| Fig. 4.3 Correlation between performance variation and objective metrics variation..... | 49 |
| Fig. 4.4 Mean values of subjective variations in two conditions by task | 52 |

NOMENCLATURE

| | |
|----------|--|
| HFE | Human Factors and Ergonomics |
| NASA | National Aeronautical and Space Administration |
| NASA-TLX | NASA Task Load Index |
| RAS | Robotic-assisted Surgery |

ABSTRACT

Author: Wu, Chuhao. MSIE

Institution: Purdue University

Degree Received: August 2019

Title: Eye tracking and EEG measures for workload and performance in robotic surgery training

Committee Chair: Denny Yu

Robotic-assisted surgery (RAS) is one of the most significant advancements in surgical techniques in the past three decades. It provides benefits of reduced infection risks and shortened recovery time over open surgery as well as improved dexterity, stereoscopic vision, and ergonomic console over laparoscopic surgery. The prevalence of RAS systems has increased over years and is expected to grow even larger. However, the major concerns of RAS are the technical difficulty and the system complexity, which can result in long learning time and impose extra cognitive workload and stress on the operating room. Human Factor and Ergonomics (HFE) perspective is critical to patient safety and relevant researches have long provided methods to improve surgical outcomes. Yet, limited studies especially using objective measurements, have been done in the RAS environment.

With advances in wearable sensing technology and data analytics, the applications of physiological measures in HFE have been ever increasing. Physiological measures are objective and real-time, free of some main limitations in subjective measures. Eye tracker as a minimally-intrusive and continuous measuring device can provide both physiological and behavioral metrics. These metrics have been found sensitive to changes in workload in various domains. Meanwhile, electroencephalography (EEG) signals capture electrical activity in the cerebral cortex and can reflect cognitive processes that are difficult to assess with other objective measures. Both techniques have the potential to help address some of the challenges in RAS.

In this study, eight RAS trainees participated in a 3-month long experiment. In total, they completed 26 robotic skills simulation sessions. In each session, participants performed up to 12 simulated RAS exercises with varying levels of difficulty. For Research Question I, correlation and mixed effect analyses were conducted to explore the relationships between eye tracking

metrics and workload. Machine learning classifiers were used to determine the sensitivity of differentiating low and high workload with eye tracking metrics. For Research Question II, two eye tracking metrics and one EEG metric were used to explain participants' performance changes between consecutive sessions. Correlation and ANOVA analyses were conducted to examine whether variations in performance had significant relationships with variations in objective metrics. Classification models were built to examine the capability of objective metrics in predicting improvement during RAS training.

In Research Question I, pupil diameter and gaze entropy distinguished between different task difficulty levels, and both metrics increased as the level of difficulty increased. Yet only gaze entropy was correlated with subjective workload measurement. The classification model achieved an average accuracy of 89.3% in predicting workload levels. In Research Question II, variations in gaze entropy and engagement index were negatively correlated with variations in task performance. Both metrics tended to decrease when performance increased. The classification model achieved an average accuracy of 68.5% in predicting improvements.

Eye tracking metrics can measure both task workload and perceived workload during simulated RAS training. It can potentially be used for real-time monitoring of workload in RAS procedure to identify task contributors to high workload and provide insights for training. When combined with EEG, the objective metrics can explain the performance changes during RAS training, and help estimate room for improvements.

1. INTRODUCTION

1.1 Robotic surgery

Minimally invasive surgery (MIS) is one of the greatest surgical innovations of the past three decades (Diana & Marescaux, 2015). It allows surgeons to view through an endoscope and manipulate the tissues or organs with thin instruments through small incisions. Compared with traditional open surgery, it offers benefits of less trauma, reduced infection risks, decreased postoperative pain, and shortened patient recovery time (Fuchs, 2002; Verhage, Hazebroek, Boone, & Van, 2009). Despite benefits, early MIS (also referred as laparoscopic surgery) has been observed to cause higher cognitive and physical workload than that of open surgery (Berguer, Smith, & Chung, 2001; Hemal, Srinivas, & Charles, 2001), due to limitations in tactile sensation, video displays, interface design, and the disconnect of separating the surgeons' hands from target organs (Ballantyne, 2002; Lowndes & Hallbeck, 2014; Yu, Lowndes, Morrow, et al., 2016). Specifically, laparoscopic surgery uses a two-dimensional vision and results in loss of depth perception to some extent; also, the camera is held by an assistant so there is a separation of vision and physical operation of the instruments, increasing the difficulty of eye-hand coordination (Supe, Kulkarni, & Supe, 2010). Studies have shown that the drawbacks of MIS can cause problems to surgeon's health and performance (Hemal et al., 2001; Marucci et al., 2000)

Advances in robotic-assisted MIS (RAS) systems have the potential to address some of the ergonomic limitations observed in MIS (Moorthy et al., 2004; Yu et al., 2017). The use of robots in surgery commenced in 1994 when the first voiced controlled camera holder prototype robot was approved by the FDA (Sackier & Wang, 1994). In 1997, Intuitive Surgical Inc. developed and marketed the da Vinci system, a master-slave manipulator that was a breakthrough in RAS (Palep, 2009). RAS provides the same benefits of MIS while eliminating many of the pitfalls in conventional laparoscopy. Potential advantages include increased dexterity, adjustable console positions, and stereoscopic visualization (Lanfranco, Castellanos, Desai, & Meyers, 2004). And the use of RAS is expected to increase with more functional systems being developed (Rassweiler et al., 2017).

Although studies have shown that RAS is less physically stressful than laparoscopy, there are still challenges in this new technology that may lead to high cognitive workload for the surgical team. Existing robotic systems (Fig. 1.1) are relatively large and cumbersome which can increase the coordination difficulty in today's already crowded operating rooms. And the physical separation between surgeons and their patients or/and surgical teams may increase their stress to maintain awareness of the environment. For example, similar to laparoscopy, flow disruptions in robotic surgery have been observed to occur frequently, and disruption severity were associated with increased self-reported workload (Blikkendaal et al., 2017; Weber et al., 2018). Other challenges are due to the unique interfaces (Fig. 1.2) and technique complexity. Surgeons need to familiarize themselves with the interfaces and operations before starting live surgeries, which may lead to long learning time (Steinberg, Merguerian, Bihrlé, & Seigne, 2008) and high workload (Catchpole et al., 2018). The lack of tactile feedback is another known disadvantage that could increase surgeon workload (Talamini, Chapman, Horgan, & Melvin, 2003; Wottawa et al., 2016) and lead to adverse surgery outcomes (Hubens, Ruppert, Balliu, & Vaneerdeweg, 2004). These new challenges necessitate additional studies on assessing the workload and training process of RAS.



Fig. 1.1 Example of RAS system components
 Surgeon console (left), Patient cart (middle), HD Vision Cart (left)
 Figure from da Vinci S System User Manual (Intuitive Surgical, Inc., 2014)



Fig. 1.2 Example of RAS surgeon console interfaces
 Stereo Viewer (left), Master controller (right), Footswitch panel (bottom)
 Figures from da Vinci S System User Manual (Intuitive Surgical, Inc., 2014)

1.2 Human factors approach to surgery

The awareness of HFE in the field of healthcare has been increasing. One of the first HFE studies on medical safety was conducted in the early 1960's, which examined medication administration errors (Safren & Chapanis, 1960). On November 29, 1999, the Institute of Medicine released the report: "To Err is Human: Building a Safer Health System", which recognized HFE and its systems approach as critical for patient safety across all healthcare domains (Kohn, Corrigan, & Donaldson, 1999). Till now, a variety of guidance documents on analyzing healthcare safety events from a HFE perspective have been published by the US Department of Health and Human Service, US Food and Drug Administration (Sawyer et al., 1996), the US Agency for Healthcare Research and Quality (Henriksen, Dayton, Keyes, Carayon, & Hughes, 2008) and other professional organizations. HFE researches can benefit both the caregiver (occupational ergonomics) and care receiver (patient safety) (Sue Hignett, Carayon, Buckle, & Catchpole, 2013). Over the past decades, there have been a number of recommendations for improving working ergonomics and reducing the risk of occupational hazards (Dawson et al., 2007; S. Hignett, 2003). These interventions range from organization changes to device design and personal well-being programs. Likewise,

substantial efforts have been invested to reduce the risk of medical errors and improve patient safety (Xie & Carayon, 2015).

Due to the critical nature of surgical interventions, surgeries account for a high proportion of medical errors, which can be translated into prolonged recovery, morbidities or mortalities (Gawande, Thomas, Zinner, & Brennan, 1999). There is a large amount of HFE literature on surgery with interests on various aspects (e.g., technical competence, non-technical skills and environment factors), and these studies have provided effective ways to improve patient outcomes and reduce surgeons' work-related injuries. For example, the physically taxing position and posture restriction during surgery has been associated with musculoskeletal injuries of surgical technicians and assistants (Davis, Fletcher, & Guillaumondegui, 2014; Sheikhzadeh, Gore, Zuckerman, & Nordin, 2009). Evidence-based studies have provided guidelines for device position and usage that can improve physical ergonomics (van Det, Meijerink, Hoff, Totté, & Pierie, 2009; Veelen, Jakimowicz, & Kazemier, 2004). Meanwhile, surgery outcomes and patient safety can be impacted by several factors. For example, poor communication or teamwork has been increasingly regarded as a causal factor for adverse surgical events (ElBardissi, Wiegmann, Henrickson, Wadhera, & Sundt, 2008; Gawande, Zinner, Studdert, & Brennan, 2003). And assessment tools for teamwork and communication have been developed to measure and improve surgeons' non-technical skills (Wahr et al., 2013). Intraoperative workload that exceeds surgeons' capacity can compromise both technical and non-technical skills and increase the chances of adverse patient outcomes (Arora et al., 2010; Wetzel et al., 2006). Therefore, substantial efforts have been put to investigate methods for measuring mental workload and stress in operating rooms (Carswell, Clarke, & Seales, 2005; Rubio, Díaz, Martín, & Puente, 2004). These measurements of communications, teamwork and workload can also be used to evaluate the effectiveness of surgical training (Moorthy, Munz, Adams, Pandey, & Darzi, 2005; Zheng, Cassera, Martinec, Spaun, & Swanström, 2010).

This research seeks to examine the opportunities for addressing RAS challenges through HFE approaches. Despite the challenges from procedure complexity, multitasking, interdisciplinary team work etc., surgeons were less likely to acknowledge the corresponding effects on their performance than other professionals (Sexton, Thomas, & Helmreich, 2001). This attitude has

discouraged applied researches especially those using subjective measures (Moorthy, Munz, Dosis, Bann, & Darzi, 2003). Hence, this research focuses primarily on objective techniques which can measure participants' behaviors and cognitive states in RAS in real-time. The intent is to compare the information derived from physiological measures or behavioral measures with more generally accepted criteria (e.g., task performance). Metrics that are proven to be relevant can provide effective assessments for RAS training.

1.3 Structure of the Document

This thesis is divided into five chapters. Chapter 2 provides the literature review of the developments and drawbacks in RAS training and objective measures applied in HFE studies, with emphasis on the surgical domain. In the beginning of Chapter 3, the framework of the study is explained, which consists of two Research Questions. The remaining part of Chapter 3 and whole Chapter 4 are devoted to describing the experimental methodology, results and discussions for the two Research Questions. Chapter 5 draws final conclusions and provides recommendations for future work.

2. LITERATURE REVIEW

2.1 Robotic Surgery Training

With advances in video imaging, endoscope technology and instrumentation, RAS has minimized the invasiveness of many surgical procedures and resulted in reduced blood loss, shortened postoperative stay and other benefits (Diana & Marescaux, 2015; Giulianotti et al., 2011; Mack, 2001). Despite the advantages, there are unique difficulties for surgeons to learn this new technique. The complexities of the system, spatial separation of bedside and console surgeons, and communication challenges impede mentored intraoperative teaching (Dulan et al., 2012a).

Curricula and instructions have been developed to address the education needs in various surgical specialties (Lee, Mucksavage, Sundaram, & McDougall, 2011), yet validations for RAS training are still insufficient compared to those for laparoscopic surgery and there is no standardized curriculum in existence for RAS (Dulan et al., 2012b; Yokoi, Chen, Desai, & Hung, 2018). Due to the development in technology, computer-based simulation training or virtual reality training has become popular because it is more cost-effective and provides a safer training environment (Bric, Lumbard, Frelich, & Gould, 2016). There are several virtual reality robotic simulators available on the market: da Vinci® Skills Simulator (dVSS) from Intuitive Surgical, Inc., Robotic Surgery Simulator (Ross) from Simulated Surgical Systems, Inc., RobotiX Mentor from 3D Systems, Inc. and Mimic dV-Trainer from Mimic Technologies, Inc., which all have been studied for validations in face, content or construct (Hung et al., 2011; Kenney, Wszolek, Gould, Libertino, & Moinzadeh, 2009; Seixas-Mikelus et al., 2010; Whittaker et al., 2015). And a comparative analysis including 105 medical students or physicians suggested that overall, da Vinci skills simulator would be the most popular choice (Tanaka et al., 2016). These simulators each provides a range of exercises for different robotic skills including instrument manipulation, camera control, and suturing. Although these have been shown to help beginners learn basic skills effectively, there is little evidence for benefiting more advanced skills (Phé et al., 2017). Another issue for simulation training is that there is little evidence for the transferability of skills gained using simulators to the real operating room (Abboudi et al., 2013; Moglia et al., 2016).

An important topic in RAS training is the learning curve. The learning curve is a graphic representation of the temporal relationship between the surgeon's mastery of a specifically assigned task and the chronological number of cases performed (Bokhari, Patel, Ramos-Valadez, Ragupathi, & Haas, 2011). It is being used to present knowledge gaining and skill improving process in surgery, and a validated learning curve contributes to the establishment of training program and facilitates the incorporation of RAS (Kaul, Shah, & Menon, 2006). The first step in the evaluation of a learning curve is to select an appropriate outcome that measures the surgeon's ability to perform a particular task (Tekkis, Senagore, Delaney, & Fazio, 2005). Several studies have identified learning curve by scoring the videos of simulated training using structured assessment tools (Chang, Satava, Pellegrini, & Sinanan, 2003; Hernandez et al., 2004). Studies on laparoscopic surgery also used patient outcomes of live clinical cases as the key measurement of learning curve (Shah, Joseph, & Haray, 2005), but this practice is scarce for RAS. All simulators provide users with an objective scoring based on criteria like timing and accuracy of task completion, which could be used to assess RAS learning curve (Brinkman et al., 2013; Lerner, Ayalew, Peine, & Sundaram, 2010).

One primary goal of studying learning curve is to define the cases or time required to achieve technical competence. Studies have shown that proficiency of robotic surgery (when the slope of the operative time curve becomes less steep) can be achieved after 20~30 cases (Foote & Valea, 2016). Yet proficiency is not the same as mastery and improvements have been observed throughout all cases (Lin, Frey, & Huang, 2014). Because of the existences of different simulator and exercises, the learning curve in simulated training is more uncertain. Yet for both live cases and simulated training, the learning curve has been found to vary greatly for individuals (Schreuder, Wolswijk, Zweemer, Schijven, & Verheijen, 2012). It has also been observed that trainees who took longer to become proficient on the simulator showed a faster skills decline (Zhang & Sumer, 2013). Therefore, it is important to provide feasible metrics other than performance provided by simulator and help determine if the training length is sufficient for proficiency. Additional metrics would also be useful for assessing the transferability of training outcomes to live surgery.

2.2 Objective measures in HFE

HFE studies have developed a number of measures for improving occupational health and patient safety in traditional open surgery and laparoscopic surgery. Subjective measures like survey and questionnaire have already been applied in RAS environment to assess surgeons' workload. For example, several studies have compared mental workload or stress in RAS and laparoscopic surgery through self-reported methods, e.g., National Aeronautics and Space Administration Task Load Index (NASA-TLX) (Lee et al., 2014), the surgery task load index (SURG-TLX) (Moore et al., 2015), Multiple Resources Questionnaire, and Dundee Stress State Questionnaire (Klein et al., 2012). These measures have been validated in previous studies and were successful in distinguishing mental workload between surgical techniques (Koca et al., 2015), team roles (Yu, Lowndes, Thiels, et al., 2016), and experience level (Klein et al., 2008). Despite advantages, subjective approaches are limited by potential bias (e.g., between subject variability and ability to self-assess cognitive capacity), disrupt the surgical task, and are only available at the completion of the case when they are typically administered (Carswell et al., 2005; Miller, 2001; Young, Brookhuis, Wickens, & Hancock, 2015). These drawbacks may be avoided by using various physiological or behavioral measures which provide objective and continuous data of operator states.

Compared with subjective measures, physiological measures are advantageous in four aspects. First of all, the assessment can be viewed as objective since it is independent of user's perception or attitude (Kivikangas et al., 2011), which increases its reliability. Many physiological measures can provide multidimensional information, which means they are sensitive to more than one cognitive process (Damos, 1991). Although physiological measures can be physically obtrusive in that they need to have direct contact with the user's body, they do not interfere with task procedure and are considered procedurally unobtrusive. Lastly, physiological measures are continuous signals which provides the possibility of real-time monitoring and interventions. With advances in wireless sensors and signal analytics, physiological measures are becoming more feasible in the operating room and provide objective approaches to continuously monitor surgeons' states without interfering intraoperatively (Dias, Ngo-Howard, Boskovski, Zenati, & Yule, 2018). Several studies have attempted to objectively measure physical workload during robotic surgery by using surface

electromyography and motion tracking (Lee et al., 2014; Yu et al., 2017; Zihni, Ohu, Cavallo, Cho, & Awad, 2014) and proved the ergonomic advantages of RAS over laparoscopy.

There are limited studies that evaluated mental workload or cognitive workload in RAS through objective measures. Yet the relationship between objective measures and mental workload has been published in many domains. Examples of physiological metrics include electroencephalogram (EEG), pupillometry, and heart rate variability (HRV). EEG can relate to several cognition processes by recording electrophysiological activity of the cerebral cortex (Lean & Shan, 2012). The electrical signals are usually processed into frequency domain in alpha, beta, gamma, delta and theta band. For example, brain wave rhythm in alpha band is prominent when subjects were asked to relax (Antonenko, Paas, Grabner, & van Gog, 2010). Instead of using a single band, studies have developed metrics using the combination of frequency bands to measure mental workload, engagement, vigilance, and fatigue (Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2014; Kamzanova, Kustubayeva, & Matthews, 2014). The application of these metrics to surgery is still nascent, but preliminary works by Guru and colleagues have showed that EEG metrics correlated with objective performance and with SURG-TLX subscales (mental and temporal demand) during robotic procedures (Guru, Esfahani, et al., 2015; Guru, Shafiei, et al., 2015). Despite the informative metrics derived from EEG signals, the intrusive setup procedure and susceptibility to motion/muscle artifacts have limited EEG's application and reliability in the fast-paced and dynamic surgical environment (Ayaz et al., 2012; Cao, Chintamani, Pandya, & Ellis, 2009; Miller, 2001). Researchers have worked towards wireless EEG systems that are more resilient to adverse environment, which may help exploit the full potential of EEG (Debener, Minow, Emkes, Gandras, & Vos, 2012).

Pupil diameter and other eye-related metrics are more widely used than before. The development of multifunctional eye tracker has addressed many usability and reliability concerns and this technology can provide both physiological and behavioral measurements. There has been a growing number of applications of eye tracking in surgical training and education (Henneman, Marquard, Fisher, & Gawlinski, 2017; Tien et al., 2014). These studies showed that expert and novice surgeons have different gaze patterns (Khan et al., 2012; Wilson et al., 2010) and projecting experts' gaze patterns to trainees could improve their performance in laparoscopic tasks and

accelerating their learning process (Chetwood et al., 2012; Wilson et al., 2011). Eye-related metrics have also shown strong associations with mental workload in many domains (Beatty, 1982; Greef, Lafeber, Oostendorp, & Lindenberg, 2009; Marquart, Cabrall, & de Winter, 2015). Preliminary works have applied several eye-related metrics to measure surgical workload. For example, peak pupil size was shown to increase with task difficulty while novices transported rubber objects over dishes with different target sizes and distances (Zheng, Jiang, & Atkins, 2015). Low blink frequency range was found to be associated with higher NASA-TLX ratings during simulated laparoscopic tasks (Zheng et al., 2012). A common limitation of these studies is the reliance on basic tasks or the focus on laparoscopic techniques. The accuracy of eye tracking measures for RAS tasks with more complex interfaces remains unknown. Research is needed to determine the impact of robotic interfaces (surgeons look through a surgical console during the surgery) and high technical complexity of RAS on eye tracking implementation and its ability to assess workload.

Compared with EEG measurement and eye tracker, heart rate sensors are easier to implement and have been extensively studied in HFE. Common applications include measurements for mental and physical workload (Meshkati, 1988; Roscoe, 1992) and fatigue (Egelund, 1982). Metrics derived from heart rate or heart rate variability (HRV) have been frequently used to assess surgeons' stress and cognitive workload (Dias et al., 2018; Rieger, Stoll, Kreuzfeld, Behrens, & Weippert, 2014). Despite the wide usage, there are debates about the reliability of some metrics. For example, emotional stimulus and physical workload can also increase heart rate (Jorna, 1992, 1993), and many studies have noted that HRV is not sensitive enough for measuring mental workload (Gabaude, Baracat, Jallais, Bonniaud, & Fort, 2012; Nickel & Nachreiner, 2003).

Other objective measures were less used due to feasibility or reliability issues. For example, facial thermography was proposed as a non-intrusive measurement for mental workload (Murai, Okazaki, Stone, & Hayashi, 2007; Or & Duffy, 2007) but this technique is difficult to apply in RAS since the surgeon's head will be in the console. Similarly, several studies have explored salivary stress hormones (e.g., cortisol) as indicators of workload (Metzenthin et al., 2009); however, salivary measurements are intrusive to measure in the operating room and can be unreliable since diurnal rhythm of cortisol secretion vary naturally throughout the day (Abdelrahman et al., 2016).

Metrics obtained from the objective measurements mentioned above can potentially identify behaviors and cognitive processes related to performance change. For example, mental workload is known to affect task performance, and several studies have found worse surgical task performance was associated with higher mental workload as measured by various physiological signals (Guru, Shafiei, et al., 2015; Zheng et al., 2015). Using workload-related physiological signals, studies also differentiated between experienced surgeons and novices (Law, Atkins, Kirkpatrick, & Lomax, 2004; Zheng et al., 2010). In addition to workload, studies have shown that EEG measures reflected changes in engagement, which were correlated with performance in vigilance test (Berka et al., 2007). And EEG measures for engagement have been used to develop adaptive automation which could improve task performance (Baldwin & Penaranda, 2012; Freeman, Mikulka, Scerbo, & Scott, 2004).

2.3 Classification analysis

As mentioned above, objective measures were widely applied for measuring operators' functional states like workload and fatigue. A useful technique of detecting functional states from continuous signals is classification analysis, or more specifically, supervised machine learning classification. The first step in applying classification is to label the functional states that need to be detected. Then relevant features need to be identified and extracted from the raw objective signals. These features will be used and train a classification model that learns to predict labels automatically, and the model will be further validated with testing data. Due to the dynamic nature of the physiological measures, conventional linear approaches are not always appropriate in modelling cognitive states and machine learning classification provides more powerful algorithms to extract information from physiological signals (Chen, Zhao, Zhang, & Zou, 2015). Another advantage of machine learning technique is that it can combine the information of multiple measures, since one measure may not be enough for reliably detecting subtle changes in cognitive states. Studies in different domains have used objective measures to detect stress (Khosrowabadi, Quek, Ang, Tung, & Heijnen, 2011; Lee, Chong, & Lee, 2017), alertness and drowsiness (Chen et al., 2015; Wang & Xu, 2016), and workload (Henelius, Hirvonen, Holm, Korpela, & Muller, 2009; Putze, Jarvis, & Schultz, 2010). In addition to cognitive states, it is also possible to classify surgeons' level of expertise through physiological measures (Richstone & Richstone, 2010).

Although the general way of applying machine learning classification is well-established (Moustafa, Luz, & Longo, 2017), there are many details where studies vary from each other. First of all, for supervised machine learning, it is required to label the data for cognitive states which need to be predicted. The most common approaches to obtain the labels is through predefined task characteristics (e.g. high/low workload task) (Halverson, Estepp, Christensen, & Monnin, 2012) or through participants' subjective reflection of their cognitive states (Stemberger, Allison, & Schnell, 2010). Based on whether the model is trained individually or collectively for all participants, classification models can be divided into individual model and population model. The advantage of training models individually is that the models can account for participants' physiology differences and even select different features (Ferreira et al., 2014; Wilson & Russell, 2003). However, the benefit of doing so is not significant and in practice a generalized model for all participants is more common. One of the most important goals of using machine learning is to achieve real-time classification. One real-time approach is to detect the state using fragments of the data, which is also called the sliding window. Studies have reported a trade-off between window size and classification accuracy: the bigger the window, the better the classification performance (Grimes, Tan, Hudson, Shenoy, & Rao, 2008; Solovey, Zec, Garcia Perez, Reimer, & Mehler, 2014). There are also multiple algorithms that are capable of classification: artificial neural networks (Baldwin & Penaranda, 2012; Wilson & Russell, 2003); random forest classifier (Rajan, Selker, & Lane, 2016; Zhou, Jung, & Chen, 2015), support vector machine (SVM) (Walter, Schmidt, Rosenstiel, Gerjets, & Bogdan, 2013), Naïve Bayes (Grimes et al., 2008); logistic regression (S. Chen, Epps, & Chen, 2013), and quadratic discriminant analysis (Ferreira et al., 2014). Yet it appeared that the choice of classifier did not make a large difference in model performance; feature generation and selection may be more important for accuracy (Solovey et al., 2014).

3. RESEARCH QUESTION I: WORKLOAD MEASURES

3.1 Research Framework

The main research question this thesis seeks to answer is: Can objective metrics used in HFE studies help address some of the RAS challenges? This main research question was further broken down into two research questions:

Research Question 1 (RQ I): Can objective metrics measure workload during RAS training?

Research Question 2 (RQ II): Can workload measure and other relevant measures explain performance improvement in RAS training?

The experiment design was based on RAS training curriculum, and data collection procedure remained consistent throughout the two RQs. Data and results from RQ I contributed to the study for RQ II. RQ I focused on measurements of mental workload during training process. As mentioned in the introduction, RAS can be cognitively demanding, yet there is a lack of studies for objectively measuring workload in RAS. Among the devices which are feasible on RAS environment, eye tracker can provide physiological and behavioral measures that are potentially responsive to changes in workload. Three questions were examined for proposed metrics:

Question 1 (Q1): Whether metrics will increase when task difficulty increase, or decrease when task difficulty increase?

Question 2 (Q2): Whether metrics will increase when subjective measures of workload increase, or decrease when subjective measures increase?

Question 3 (Q3): What is the performance of predicting level of subjective workload with eye tracking metrics?

3.2 Methodology

3.2.1 Robotic system and tasks

The Da Vinci Skill Simulator (dVSS, Intuitive Surgical, Inc. Sunnyvale, CA) was used to produce experimental tasks and train participants for RAS. The system consisted of a surgeon console with controls (e.g., foot pedals, master controls, and controls to adjust positioning) and tele-surgical robotic arms. The console also included a widely-used simulation software (M-Sim[®]) provided by

the dVSS manufacturer, which enabled trainees to perform simulated exercises without physically activating the actual robotic arms. Both the console and the software were used in this study.

Tasks and difficulties were selected from the simulation software and recommendations from the surgical education community (Alzahrani et al., 2013; Perrenot et al., 2012). These tasks required trainees to use fundamental RAS skills like camera control, endowrist manipulation, clutching, needle control, and needle driving to transfer or suture objects (as shown in Table 3.1). Depending on the specific task, 1-3 levels of difficulty were available in the simulation software, and all levels were used in the study. A task at a certain level is referred as an *exercise* in this paper. Tasks analysis based on human processor model (Card, Moran, & Newell, 1986; Feyen & Liu, 2001) and Therbligs (Gilbreth & Kent, 1911) was conducted to describe the task demands across task levels. See Table A.1 in Appendix for detailed task descriptions and task demands. Task order was not randomized due to the curriculum nature of the training, i.e. simpler tasks were prerequisites of more advanced tasks. Based on the task orders used in previous studies (Finnegan, Meraney, Staff, & Shichman, 2012; Kenney et al., 2009), tasks were performed in the following order: Camera Targeting, Peg Board, Ring and Rail, Sponge Suturing, Dots and Needles, and Tubes; and in each task, lower (easier) levels were presented before higher (more difficult) levels.

Table 3.1 Training tasks screenshots

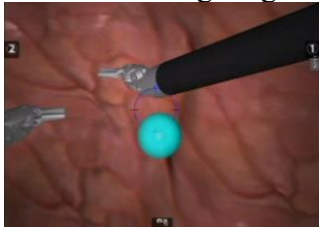


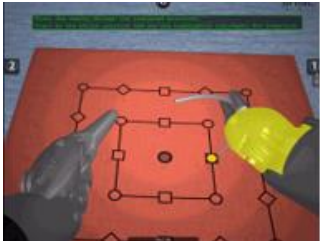


| Screenshots | Description | Screenshots | Description |
|---|---|--|---|
| <p>Camera Targeting</p>  | <p>Focus the camera on different blue spheres spread across a broad pelvic cavity. Two difficulty levels.</p> | <p>Suture Sponge</p>  | <p>Drive the needle through random targets on a deformable structure. Three difficulty levels</p> |
| <p>Peg Board</p>  | <p>Grasp rings on a stand with the left hand and pass them to the right hand before place them on a peg. Two difficulty levels.</p> | <p>Dots and Needles</p>  | <p>Insert a needle through several pairs of targets that have various spatial positions. Two difficulty levels.</p> |

Table 3.1 Continued

| Screenshots | Description | Screenshots | Description |
|---|---|--|---|
|  | Move a ring along a twisted metal rod without applying excessive force to either the ring or the rail. Two difficulty levels. |  | Drive needle through fixed targets on a cylindrical deformable structure. One difficulty level. |

3.2.2 Measurement Metrics

3.2.2.1 Performance Metric

The simulation software automatically assessed trainees' performance based on several criteria, e.g., time, economy of motion, drops, instrument collisions, excessive instrument force, instruments out of view, and master workspace range, which was summarized as an overall score (0-100%) with higher scores representing better performance (Fig. 3.1). The details for calculating this score is proprietary and have not been publicized. This overall score was recorded and used as the only measurement of performance. Due to the design of the software, this overall score was instantly displayed upon completion of each exercise and the participant saw their performance score.

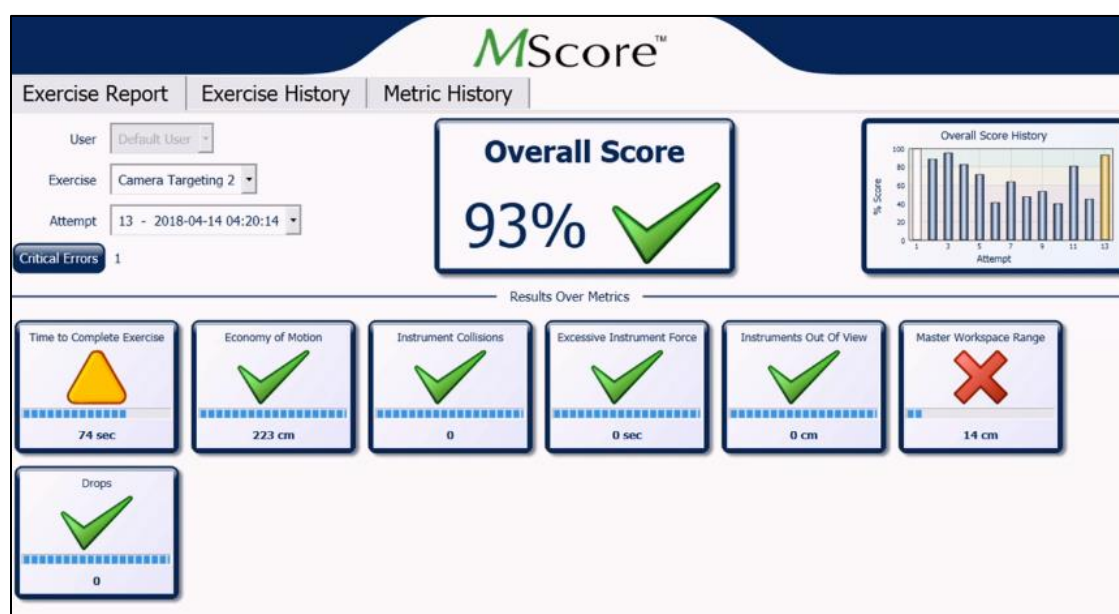


Fig. 3.1 Screenshot of task performance

3.2.2.2 Subjective Metric

The NASA-TLX survey (Hart & Staveland, 1988) was used to assess perceived workload of participants. The NASA-TLX contains six sub-dimensions of workload (mental demand, physical demand, temporal demand, performance, effort, and frustration) and each was rated on a visual analogue scale that ranged from 0 (very low) to 10 (very high) (see APPENDIX B). Scores from each dimension were added up to calculate the final NASA-TLX workload score, resulting in a value of 0 to 60. Although a weighted NASA-TLX has also been used by other investigators, many studies have demonstrated a summed score (referred to as Raw TLX) as an acceptable implementation of the NASA-TLX tool (Hart, 2006).

3.2.2.3 Eye Tracking Metrics

A wearable wireless eye tracking system, Tobii Pro Glasses 2.0, (Tobii Technology AB, Danderyd, Sweden) was used to binocularly sample eye movement at 50Hz. The eye tracker consisted of two major parts: a camera was located in the middle of the glasses frame (outer side) to record the view of scene while sensors were mounted in the inner side of the glasses frame to capture eye movements and pupil diameter. Recordings were annotated using the Tobii Pro Lab Software (Tobii Technology AB, Danderyd, Sweden) and extracted for further analysis. Four metrics can be calculated based on eye tracker data: pupil diameter, gaze entropy, fixation duration, and PERCLOS (Percentage of eyelid closure). Even though pupil diameter is a pupillometry metric, for simplicity reason, all of the four metrics can be referred to as eye tracking metrics in this study.

Pupil diameter: This metric was estimated by the eye tracking system using images of the eyes. Previous work showed association between pupillary dilations and increased cognitive load (Beatty, 1982; Beatty & Kahneman, 1966). Pupil diameter of left and right eyes was averaged as one measurement.

Gaze entropy: An index that measured visual scanning randomness and was used as a measure of mental workload in aviation tasks (Harris, Tole, Stephens, & Ephrath, 1981; Tole, 1983). It takes the distribution of all gaze points and summarizes the probability of gaze falling on each position. Therefore, gaze entropy tends to be lower when gaze points are in proximity to each other. It was calculated based on Shannon entropy theory (Leandro L. Di Stasi et al., 2016; Shannon, 2001) :

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

where $p(x, y)$ was the probability of gaze falling in the (x, y) position. 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 was monitored during the exercise, across all possible x and y in the field of view.

Fixation duration: The total amount of time spent in fixations. Studies have suggested that fixation duration reflected high information processing load and increased as workload increased (Greef et al., 2009; Morris, Rayner, & Pollatsek, 1990a; Recarte & Nunes, 2000). We scaled the time duration to the percentage of time in the exercise duration:

$$FD_{\%} = \frac{\text{Sum of fixation durations}}{\text{Exercise duration}} \times 100\%.$$

Percentage of eyelid closure (PERCLOS): PERCLOS was usually calculated as the percentage of time during which the pupils were covered by the eyelids by more than 80% of their area (Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994). Studies showed that higher PERCLOS reflected increased fatigue and decreased vigilance (Marquart et al., 2015; Singh, Bhatia, & Kaur, 2011; Sommer & Golz, 2010). It has also been used as a machine learning feature to predict workload (Halverson et al., 2012; Tian, Zhang, Wang, Yan, & Chen, 2019). In this study, since the device did not support eyelid closure measurement, it was estimated by the percentage of time duration (per exercise) where neither left pupil or right pupil is detected. Since participants' head movements were constrained, this estimation will not be confounded by participants looking away. It can be potentially confounded by missing data (lost pupil frames due to device malfunction), which was 1% for our device.

3.2.3 Study procedure

3.2.3.1 Participants

This study was reviewed by the university's Institutional Review Board. Study population was surgical trainees who needed robotic skills training (i.e., limited previous robotic experience). Eight surgical trainees from a large academic medical school were recruited voluntarily. All of the participants were right-hand dominant, 4 were female, and the mean (\pm standard deviation) of age was 26 ± 1.6 years. None had prior clinical RAS experience.

3.2.3.2 Procedure

The experiment procedure was based on RAS training curricula. The participants attended training sessions periodically, and they were asked to complete the same 6 tasks (12 exercise) in each session. Sessions were scheduled based on robotic system availability. Participants were informed of the session schedule at least one week in advance. Data collection was conducted when any participant confirmed attendance.

For each session, 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 system. The system was calibrated at the beginning of each session. Baseline pupil diameter for the participants was collected following procedures recommended by previous works (Beatty & Lucero-Wagoner, 2000; Marshall, 2000; Mosaly, Mazur, & Marks, 2017). Specifically, each participant looked at the center of a white screen for 10 seconds (minimum diameter) and then a black screen (maximum diameter) for 10 seconds.

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. 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 all 12 exercises. To maintain consistency with the trainees' curriculum and system usage schedule, the time constraint of each session was 45 minutes. Therefore, considering participants' skill and capability, some advanced difficulty levels were not completed in early phase of training. After completing each exercise, the participant completed a NASA-TLX survey. Eye tracking data was continuously recorded throughout the session. Fig. 3.2 shows the examples of participants wearing eye tracker and performing tasks on the simulator.



Fig. 3.2 Participants wearing the eye tracker (left) and performing tasks (right)

3.2.4 Data Analysis

3.2.4.1 Data scaling

Pupil diameter and gaze entropy were normalized using the feature scaling (Jayalakshmi & Santhakumaran, 2011) to scale the data to the range of [0,1], accounting for potential variation from individual difference in baseline pupil size and facilitating the comparison between different variables. It also prevents the distortion in analysis caused by variable magnitude difference (Al Shalabi, Shaaban, & Kasasbeh, 2006). For each participant, the maximum value and minimum value from all sessions were used to scale his/her metrics, as shown in the formula below:

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

3.2.4.2 Workload labeling

In order to examine Q3: What is the performance of predicting level of subjective workload with eye tracking metrics? Workload levels were determined by categorizing the total NASA-TLX scores into either high or low workload. Although there is still much debate on what NASA-TLX threshold is considered “high workload,” some studies observed that scores over 50-55 (out of 100) may lead to increased errors (Colle & Reid, 2005; Mazur et al., 2014; Mazur, Mosaly, Hoyle, Jones, & Marks, 2013; Yu, Lowndes, Thiels, et al., 2016). Therefore, in this study, scores above 30 (out of 60) were categorized as high workload. For low workload, we assumed that the workload scores

were normally distributed, and the number of low workload observations should be the same as those in the high end. Scores in the middle were not used for classification considering that they were ambiguous and may not necessarily represent either high or low workload.

3.2.4.3 Statistical Analysis

Three different analysis techniques were used to examine the 3 questions. Significance level for all analyses was set at $\alpha = 0.05$. When appropriate, p-values were corrected using Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

Q1: Mixed effects models were used to determine eye tracking metrics' sensitivity to changes in task levels (difficulty). This approach accounted for random effects of subject and repeated measures by allowing varying intercept for each subject (Cnaan, Laird, & Slasor, 1997; Dingemanse & Doehrmann, 2013). Each task was fitted by separate models, resulting in 5 models (Task Tube had only one level of difficulty and was not tested).

Q2: Repeated measure correlation: r_{rm} (Bakdash & Marusich, 2017), were used to examine relationships between eye tracking metrics and NASA-TLX ratings. Instead of the more common Pearson correlation, r_{rm} coefficient was estimated using analysis of covariance, where participant was treated as a factor level. This technique gives a more accurate estimation of the association between two variables when underlying individual factors can affect the relationship. The formula of r_{rm} is expressed in the form of sum of squares:

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

where SS_{Error} is the residual sum of squares of the linear model: $Measure\ 1 = \beta_0 + \beta_1 \times Participants + \beta_2 \times Measure\ 2 + \epsilon$; $SS_{Measure} + SS_{Error}$ is the residual sum of squares of the linear model: $Measure\ 1 = \beta_0 + \beta_1 \times Participants + \epsilon$.

Q3: Machine learning algorithms were used to explore the joint capability of various eye tracking features for detecting high workload. Three different algorithms were used: logistic regression, Naïve Bayes algorithm and Support Vector Machine (SVM), all of which have been used for workload classification in previous studies (So, Wong, Mak, & Chan, 2017; Solovey et al., 2014).

For SVM, three kernel specifications were used: Linear, Gaussian and Polynomial. Details for the three algorithms can be found in an introductory book for statistical learning (James et al., 2013). For example, Naïve Bayes was based on Bayesian theorem: $P(C_j|X) \propto P(C_j) \prod P(x_i|C_j)$: the probability of a certain class, given all evidence, was the product of prior probability of the class and all conditional probability of evidences. And its main advantages were the effectiveness for small datasets (Jyothi & Bhargavi, 2009) and applicability to different types of data (Domingos & Pazzani, 1997). A k-fold cross validation procedure was used for model training and testing (Hastie, Friedman, & Tibshirani, 2001). Based on sample size, three folds were performed. A confusion matrix (Fawcett, 2006) was used to determine the accuracy and sensitivity of eye metrics in predicting workload.

3.3 Results

3.3.1 Descriptive data

RQ I spanned about 1.5 months. A total of 15 sessions across all participants were collected over the study period. Two participants attended 3 sessions, 3 participants attended 2 sessions, and 3 participants attended 1 session. A total of 168 exercises were collected, including performance scores, NASA-TLX ratings, and eye tracking features. Minimum exercises completed in a session was $n = 8$, and all participants completed each exercise at least once. Some participants did not complete all 12 exercises as explained in the methods. Average and standard deviation of exercise completion time was $194s \pm 157s$. The standard deviation was large because difficult exercises took significantly more time than easy exercises. Mean value of all measurements were reported in Table 3.2. The trend of task performance and NASA-TLX score in different difficulty level is shown in

Fig. 3.3. Histograms of NASA-TLX and subscales can be found in APPENDIX C.

Table 3.2 Mean value of all metrics across tasks and levels

| Task | CT | | PB | | RR | | SS | | | DN | | T |
|-----------------------------|------|------|------|------|------|------|------|------|------|------|------|------|
| Level | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 3 | 1 | 2 | 1 |
| Performance metric | | | | | | | | | | | | |
| Performance | 77.6 | 69.9 | 77.2 | 89.7 | 88.9 | 65.0 | 68.8 | 68.8 | 64.2 | 77.6 | 68.4 | 56.7 |
| Subjective metric | | | | | | | | | | | | |
| NASA-TLX | 13.6 | 19.7 | 15.6 | 17.0 | 17.3 | 30.4 | 24.4 | 26.1 | 26.6 | 25.9 | 26.8 | 30.8 |
| Eye tracking metrics | | | | | | | | | | | | |
| Pupil diameter | 0.54 | 0.62 | 0.67 | 0.71 | 0.47 | 0.59 | 0.45 | 0.54 | 0.53 | 0.48 | 0.54 | 0.63 |
| Gaze entropy | 0.50 | 0.60 | 0.58 | 0.63 | 0.38 | 0.76 | 0.51 | 0.68 | 0.70 | 0.60 | 0.60 | 0.72 |
| Fixation duration | 0.84 | 0.80 | 0.81 | 0.79 | 0.83 | 0.83 | 0.81 | 0.84 | 0.84 | 0.87 | 0.84 | 0.81 |
| PERCLOS | 0.07 | 0.09 | 0.10 | 0.11 | 0.10 | 0.08 | 0.13 | 0.08 | 0.08 | 0.08 | 0.10 | 0.12 |

* CT: Camera Targeting, PB: Peg Board, RR: Ring and Rail, SS: Suture Sponge, DN: Dots and Needles, T: Tubes

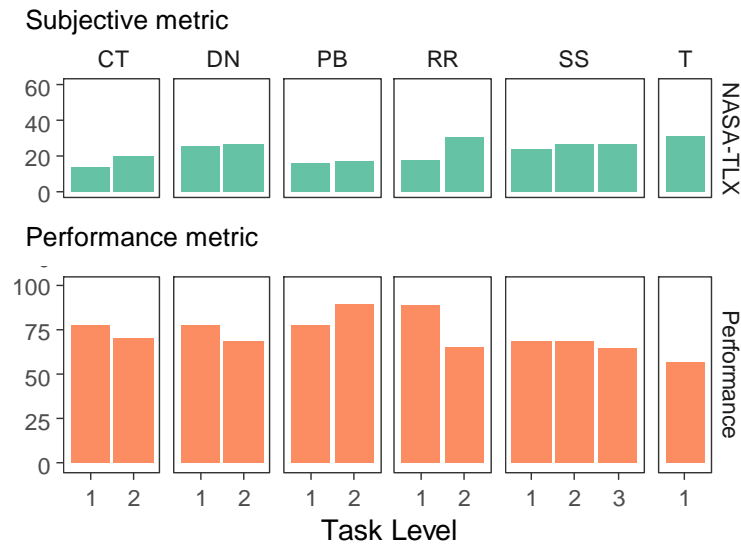


Fig. 3.3 Mean value of task performance and NASA-TLX across tasks and levels

3.3.2 Eye Tracking Metrics and Task Difficulty

Mixed effect models were used to test eye tracking metrics' sensitivity to changes in task difficulty. With task goal and skill remaining consistent, the simulator increased difficulty levels by incorporating additional task requirements, which increased the task load. Since changes in difficulty levels varied by task, each task was fitted with a model separately. Results for mixed

effect models are shown in Table 3.3, excluding results for fixation duration and PERCLOS, which did not reach statistical significance.

Table 3.3 Mixed models summary for effects of task level on eye tracking metrics

| | | Task | CT | PB | RR | SS | | DN |
|----------------------------|-------------|-------|--------|------|--------|--------|--------|------|
| | | Level | 2 | 2 | 2 | 2 | 3 | 2 |
| Subjective metric | | | | | | | | |
| NASA-TLX | Coefficient | | 6.13 | 1.40 | 13.13 | 3.19 | 3.65 | 2.26 |
| | <i>p</i> | | .002 | .348 | < .001 | .139 | .091 | .398 |
| | Cohen's d | | 1.58 | .42 | 2.76 | .53 | .61 | .41 |
| Eye tracking metric | | | | | | | | |
| Pupil diameter | Coefficient | | .08 | .03 | .12 | .08 | .07 | .05 |
| | <i>p</i> | | < .001 | .026 | < .001 | < .001 | < .001 | .024 |
| | Cohen's d | | 2.38 | 1.04 | 3.02 | 1.49 | 1.40 | 1.20 |
| Gaze entropy | Coefficient | | .11 | .05 | .38 | .17 | .18 | .03 |
| | <i>p</i> | | .004 | .082 | < .001 | < .001 | < .001 | .427 |
| | Cohen's d | | 1.40 | .79 | 3.91 | 1.86 | 2.01 | .41 |

* CT: Camera Targeting, PB: Peg Board, RR: Ring and Rail, SS: Suture Sponge, DN: Dots and Needles, T: Tubes

* Level 1 was the reference level

* Effect size of Cohen's d: Small - .20, Medium - .50, Large - .80, Very large - 1.20 (Sawilowsky, 2009)

The average NASA-TLX in higher level of difficulty was always higher (Table 3.2). Yet this relationship was statistically significant for only 2 of the tasks. Therefore, increase in task load did not necessarily increase subjective workload. Increasing difficulty was observed to significantly increase pupil diameter for all tasks (all *p*-values < .05). The positive coefficients suggested that pupil diameters in level 2 for all tasks were larger than that in level 1. Level effects (Cohen's *d*) in tasks were very large except for task Peg Board. However, when there were 3 levels of difficulty (Suture Sponge), post hoc analysis of Tukey test suggested that there was no difference between level 2 and 3 (*p*-value = .964).

For gaze entropy, significant effect of difficulty level was observed in the following tasks: Camera Targeting, Ring and Rail and Suture Sponge. Based on Cohen's *d*, effects were large in all of the 3 tasks. The positive coefficients suggested that gaze entropy in level 2 was greater than that of level 1. Gaze entropy between level 2 and level 3 in task Suture Sponge was not significantly different. Mean values of all eye metrics in each task level are presented in Fig. 3.4.

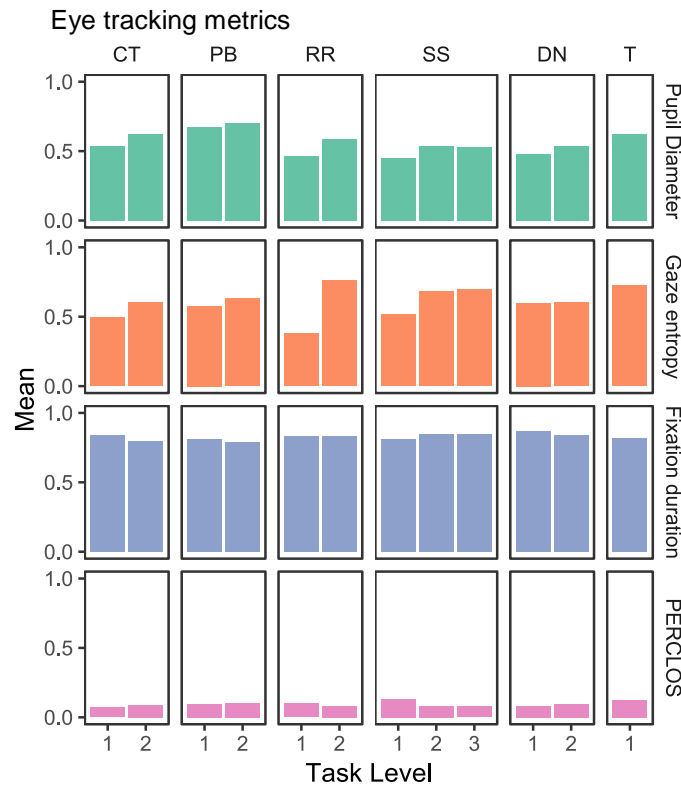


Fig. 3.4 Mean value of eye tracking metrics across tasks and levels

3.3.3 Eye tracking Metrics and NASA-TLX

NASA-TLX survey captured the perceived workload (subjective workload) from participants. And the repeated measures correlations examined whether the information aligned with that from the eye metrics. Of the four eye tracking metrics, only gaze entropy had significant correlation with NASA-TLX ratings ($r_{rm} = .51, p < .001$), indicating increase in gaze entropy with increased perceived workload. Fig. 3.5 illustrates the distribution of eye tracking measures and workload, colored by participant. The correlations also varied among tasks and all correlation values including other eye tracking metrics are reported in Table 3.4. Correlations between eye metrics and subscales of NASA-TLX can be found in APPENDIX D.

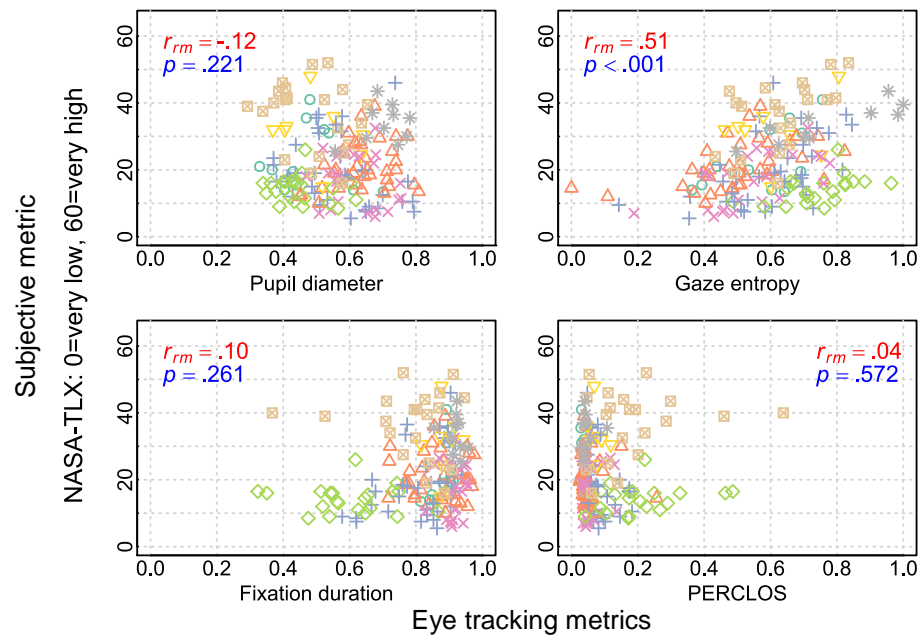


Fig. 3.5 Distribution of eye tracking metrics over workload (colored and shaped by participants)

Table 3.4 Repeated correlation between subjective metric and eye tracking metrics

| Eye tracking metrics | By Task | | | | | | | All Task n = 168 |
|----------------------|----------|--------|--------|--------|--------|--------|--------|---------------------|
| | | CT | PB | RR | SS | DN | T | |
| | | n = 30 | n = 30 | n = 30 | n = 40 | n = 26 | n = 12 | |
| Pupil diameter | r_{rm} | .52 | .19 | .58 | .43 | .55 | .63 | -.12 |
| | p | .032 | .538 | .014 | .032 | .039 | .250 | .221 |
| Gaze entropy | r_{rm} | .62 | .34 | .76 | .49 | .45 | -.42 | .51 |
| | p | .009 | .224 | < .001 | .014 | .119 | .522 | < .001 |
| Fixation duration | r_{rm} | -.20 | -.53 | -.11 | -.03 | .07 | .36 | .10 |
| | p | .522 | .032 | .736 | .851 | .815 | .561 | .261 |
| PERCLOS | r_{rm} | .20 | .70 | .13 | -.04 | -.08 | -.61 | .04 |
| | p | .522 | .002 | .702 | .851 | .815 | .263 | .572 |

* CT: Camera Targeting, PB: Peg Board, RR: Ring and Rail, SS: Suture Sponge, DN: Dots and Needles, T: Tubes

3.3.4 Classification of workload

There were 43 high workload observations with NASA-TLX scores above 30 (which is the 25% quantile). The same number of observations (43) in the lowest end were regarded as low workload, which had values below or equal to 14.5 (which is the 75% quantile). The 86 observations were partitioned into 3 sets with the size of 28, 28 and 30 for the 3-fold cross-validation. Using the machine learning classification, nine features were included to classify low/high workload: two demographic features (participant gender and trainee level (medical student/surgical resident)) and 7 eye tracking features (left/right pupil diameter mean, left/right pupil diameter standard deviation, gaze entropy, fixation duration and PERCLOS). Mean values of four main eye metrics in low workload and high workload conditions are illustrated in Fig. 3.6. Three algorithms: Logistic regression, SVM (with three different kernels) and Naïve Bayes were applied to build and validate the classification model. The average accuracy of eye tracking measures in predicting workload was 89.3% and average F1 score was 0.89. The confusion matrix of the 3-fold cross validations for each algorithm are presented in Table 3.5 (1)-(5). SVM using polynomial kernel showed the best performance, with an average accuracy of 94.3% and average F1 score of 0.94.

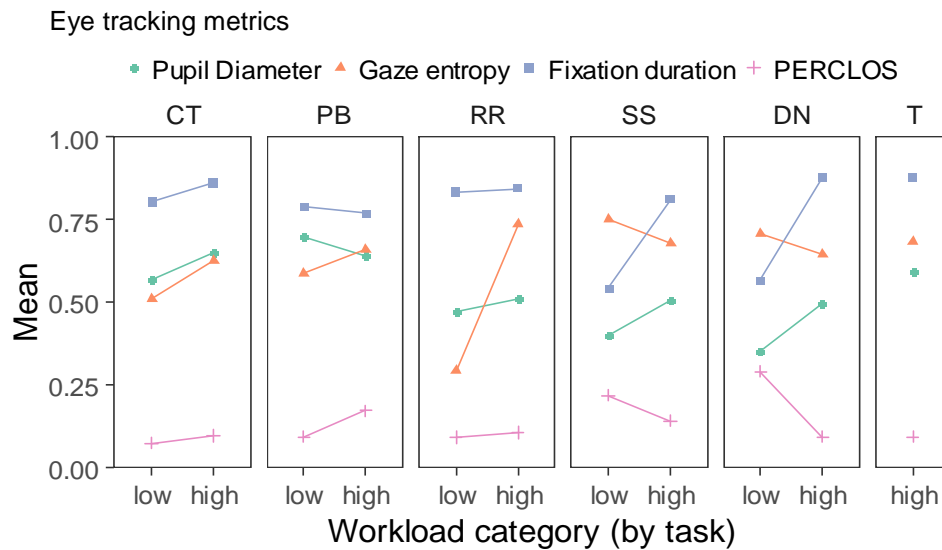


Fig. 3.6 Mean value of eye tracking metrics in high/low workload

Table 3.5 Workload classification confusion matrix for testing dataset
(86 bservations stratified into 3 folds)

(1) Logistic regression

| Logistic | | Actual class | | | |
|-----------------|---------------|----------------|---------------|-------------|-----------|
| Predicted class | High workload | High workload | Low workload | 86.3%±1.5% | |
| | | 43.0%±0.3% | 5.6%±6.9% | Precision | |
| | Low workload | 7.0%±0.3% | 44.4%±6.9% | 89.6%±12.0% | |
| | | False negative | True negative | NPV | |
| | | 86.0%±0.5% | 86.0%±0.5% | 87.4%±6.6% | 0.87±0.08 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

(2) Naïve Bayes

| Naïve Bayes | | Actual class | | | |
|-----------------|---------------|----------------|---------------|------------|-----------|
| Predicted class | High workload | High workload | Low workload | 82.8%±9.5% | |
| | | 44.1%±2.2% | 9.4%±5.6% | Precision | |
| | Low workload | 5.9%±2.2% | 40.6%±5.5% | 87.1%±5.4% | |
| | | False negative | True negative | NPV | |
| | | 88.3%±4.4% | 81.1%±11.2% | 84.7%±7.7% | 0.85±0.07 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

(3) SVM (Linear Kernel)

| SVM-Linear | | Actual class | | | |
|-----------------|---------------|----------------|---------------|-------------|-----------|
| Predicted class | High workload | High workload | Low workload | 89.1%±2.8% | |
| | | 44.1%±2.2% | 3.3%±5.8% | Precision | |
| | Low workload | 5.9%±2.2% | 46.7%±5.8% | 94.1%±10.2% | |
| | | False negative | True negative | NPV | |
| | | 88.3%±4.4% | 88.3%±4.4% | 90.8%±3.6% | 0.91±0.04 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

(4) SVM (Gaussian Kernel)

| SVM-Gaussian | | Actual class | | | |
|-----------------|---------------|----------------|---------------|-------------|-----------|
| Predicted class | High workload | High workload | Low workload | 87.4%±5.0% | |
| | | 42.9%±3.7% | 3.3%±5.8% | Precision | |
| | Low workload | 7.1%±3.7% | 46.7%±5.8% | 94.1%±10.2% | |
| | | False negative | True negative | NPV | |
| | | 85.9%±7.4% | 85.9%±7.4% | 89.6%±3.1% | 0.90±0.04 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

Table 3.5 Continued

(5) SVM (Polynomial kernel)

| SVM-Poly | | Actual class | | | |
|-----------------|---------------|----------------|----------------|------------|-----------|
| Predicted class | High workload | High workload | Low workload | | |
| | | 47.6%±4.1% | 3.3%±5.8% | 95.8%±7.2% | |
| | Low workload | True positive | False positive | Precision | |
| | | 2.4%±4.1% | 46.7%±5.8% | 94.4%±9.6% | |
| | | False negative | True negative | NPV | |
| | | 95.2%±8.2% | 95.2%±8.2% | 94.3%±5.2% | 0.94±0.06 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

Note:

The confusion matrix contains 10 indexes:

True Positive (TP): Proportion of observations that were classified correctly as high workload

False Positive (FP): Proportion of observations that were classified incorrectly as high workload

True negative (TN): Proportion of observations that were classified correctly as low workload

False negative (FN): Proportion of observations that were classified incorrectly as low workload

$$\text{Precision} = \frac{TP}{TP+FP} \times 100\%$$

$$\text{Negative predictive value (NPV)} = \frac{TN}{TN+FN} \times 100\%$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\%$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100\%$$

$$\text{Accuracy} = \frac{TP+TN}{P+N} \times 100\%$$

$$\text{F1 score} = \frac{2TP}{TP+FP+FN}$$

Results were reported as the mean ± standard deviation of 3 validations.

3.4 Discussion

This study in RQ I investigated the relationship between eye tracking measures and workload in RAS. The first question was whether eye tracking metrics can distinguish between varying work demands due to task difficulty level. The findings generally supported the sensitivity of eye tracking metrics for distinguishing the differences. Mixed effect models found significant task difficulty effects on pupil diameter and gaze entropy. The second question was whether eye tracking metrics are correlated with subjective workload and only gaze entropy was proven to be

correlated. The third question examined the feasibility of predicting subjective workload level with eye tracking metrics, and classification models have achieved high accuracy.

For Q1 and Q2, both fixation and PERCLOS showed no significant associations with workload. Previous studies have explained that fixation duration reflects information processing load (Morris, Rayner, & Pollatsek, 1990b; Recarte & Nunes, 2000), which can be a driving factor for workload. And fixation duration has been found responsive to difficulty in task that requires information acquisition from text or image (Di Stasi, Antolí, & Cañas, 2013; He, Wang, Gao, & Chen, 2012). However, for the RAS training tasks, the information processing load was fairly constant, especially after the trainees understood the goal of a task. Changes in task difficulty were attributed more to the requirements of eye-hand coordination and action planning. PERCLOS was more commonly linked to fatigue, yet also suggested as a measure for estimating workload (Halverson et al., 2012; Tian et al., 2019). When under prolonged states of low workload, it is possible that the state of drowsiness can co-occur with a state of low attentional arousal. In this robotic training setting where participants were actively engaged, low arousal levels were unlikely, which explains the low PERCLOS values observed. The low PERCLOS values also suggested that RQ II would not need to include measurements for fatigue.

Results for gaze entropy from both Q1 and Q2 supported the hypothesis that visual exploration becomes less fixed (i.e., the gaze pattern becomes more random) during more complex tasks. No previous studies have studied how workload impacts gaze entropy in robotic surgery, yet Di Stasi et al., (2016, 2017) showed that gaze entropy increased with increase in laparoscopic surgical task complexity. They have explained that without knowing the optimal visual exploration strategy, surgeons might follow a suboptimal approach, which caused gaze to move constantly, especially during complex tasks. For example, when trainees are unfamiliar with the environment and insensitive to the visual input, they cannot adopt the optimal scanning strategy. As the task difficulty increases, they need more glances to compensate the sub-optimal strategy. Similarly, when trainees are novice in console operations, they tend to make mistakes and need more movements to complete tasks. However, the relationship between gaze entropy and workload may be dependent on task structure. For RAS exercises used in this study, higher level of difficulty will inevitably require more gaze points. Although this change in task structure co-occurred with

increase in task load, it may not always be the case, especially when visual search load is low in the task.

Pupil diameter was observed to be larger in higher level of difficulty, which agrees with previous studies in surgical laparoscopy (Zheng et al., 2012) and other domains (Beatty & Kahneman, 1966; Palinko, Kun, Shyrokov, & Heeman, 2010; Schwalm, Keinath, & Zimmer, 2008). However, studies have also noted that the pupillary response to task difficulty converged with NASA-TLX rating (Marandi, Samani, & Madeleine, 2018; Recarte, Pérez, Conchillo, & Nunes, 2008). Yet our results for Q3 showed otherwise. We have also analyzed the correlations between pupil diameter and the six subscales of NASA-TLX, yet again no significant correlation has been found. It has been well established that pupil dilates during mental activities (Beatty, 1982), but workload is not the only cause to mental activities. Other factors known to revoke pupillary changes include fatigue (Morad, Lemberg, Yofe, & Dagan, 2000), emotion arousal (Partala & Surakka, 2003), and visual stimulus (Barbur, Harlow, & Sahraie, 1992). The RAS task environment abounds with visual stimulus like color changes, light changes and moving objects which could have confounded the workload's effect on pupil dilation. For example, participants who experienced high workload tended to make mistakes and see more of objects flashing or moving, which can lead to pupillary constriction. The fact that fatigue can increase pupillary variability and that both negative and positive emotions can lead to increased pupil size could further affect the consistency of pupillary measures for workload.

The relationship between NASA-TLX ratings and objective measures has been long studied, yet it remains debatable which one is a better measurement of workload. For perceived workload, NASA-TLX has been more widely used and recommended as a practical and accurate way for measurement (Carswell et al., 2005; Dias et al., 2018). Recent work by Matthews, Reinerman-Jones, Barber, and Abich (2015) found that many physiological measures as well as NASA-TLX ratings were sensitive to changes in workload, but their estimates were uncorrelated. They suggested that this was caused by individual differences or the failure on assuming workload as a unitary latent construct. Other studies explained that physiological methods gave more information on how individuals responded to workload instead of what was imposed on them (Cain, 2007; Najmedin Meshkati, Hancock, Rahimi, & Dawes, 1995). Our results of gaze entropy support the

assumption that a latent workload construct can be estimated by both subjective and objective measures. However, there was still variability between gaze entropy and NASA-TLX, which supports the argument that workload is multi-factorial and each method measured unique information. Therefore, the machine learning classification approach was used to combine four eye metrics and investigate if they can estimate the same level of workload as the NASA-TLX does but in a less disruptive way.

In the classification models, the 9 features classified between low and high workload labels with an average accuracy of 89.3%. Similar classification study using eye-related measures reported an accuracy range of 16-98% in different models (Halverson et al., 2012). In Halverson's study, there were two tasks: high workload and low workload, where participants needed to monitor more vehicles in the high workload task. In contrast, we did not classify the different tasks, but the different levels of perceived workload of participants using their NASA-TLX ratings. This method reflects more of the trainees' capacity in dealing with task demand. Classification of workload is clinically helpful to surgical education as the technique is able to provide real-time feedback on trainees' workload status, and the instances of high workload, which indicate when trainees are experiencing difficulty.

Quantitative eye metrics provide feedback regarding when the trainees' visual behaviors are inefficient and when they experience high workload. Instructors can personalize training tasks to help trainees learn how to process visual cues and practice specific skills before proceeding to more complex tasks. This study is done in a global level for tasks, i.e. we did not quantify the workload variations within an exercise. Instantaneous self-assessment may be considered for verifying workload changes within an exercise (Tattersall & Foord, 1996). Future work may also consider techniques like Hidden Markov Model for identifying high-level tasks (Lalys, Riffaud, Bouget, & Jannin, 2012) and decompose tasks and skills (Reiley & Hager, 2009), which can contribute to the understanding the relationship between workload and task structure.

4. RESEARCH QUESTION II: PERFORMANCE MEASURES

4.1 Research Framework

The research question in RQ II is: Can workload measure and other relevant measures explain performance improvement in RAS training? The main interest was individuals' performance changes between sessions, instead of inter-participant differences. Task performance is a multifaceted result driven by various factors including workload. Results from RQ I suggest that pupil diameter and gaze entropy could measure workload. Therefore, RQ II used pupil diameter as a measure of workload and gaze entropy as a measure of gaze pattern (what the metric is truly measuring). In addition, studies have shown that task engagement was related to performance and could be measured through EEG signals (Freeman, Mikulka, Scerbo, Prinzel, & Clouatre, 2000). Therefore, RQ II investigated how workload, gaze pattern and engagement changed with training. Three specific questions were examined for these objective metrics:

- Q1: For circumstances where performance improves, versus those in which performance does not improve, whether changes in the objective metrics are different?
- Q2: Whether the changes in performance between sessions are correlated with changes in objective metrics?
- Q3: What is the performance of predicting improvement with changes in objective metrics?

RQ II is a continuing extension of RQ I. The training tasks and training procedure were exactly the same and will not be repeatedly described in the methodology.

4.2 Methodology

4.2.1 Measurement Metrics

Based on results from RQ I, pupil diameter was used as a metric for mental workload and gaze entropy was used as a metric for scanning pattern. Calculation for these two metrics can be found in the methodology of RQ I.

The third objective metric was obtained from EEG measures. EEG signals were collected through the mobile EEG device EMOTIV EPOC and EMOTIV Pro software. Signals were sampled at 256

Hz on 14 channels: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8 and AF4. EEG signals were filtered using a lower-pass filter of 50 Hz and a high-pass filter of 1 Hz. Then Fast Fourier Transform was applied to transform the signals into power spectral density. Three frequency bands were extracted: theta (4-8 Hz), alpha (8-12 Hz), beta (12-25 Hz). Average EEG power of the 14 channels in different bands were combined to measure engagement during task.

Engagement refers to the ability to maintain focused attention and to remain alert to stimuli for certain periods of time (See, Howe, Warm, & Dember, 1995). It has been found predictive of learning outcomes in previous works (Chaouachi & Frasson, 2010; Ciret Galán & Beal, 2012). In this study, engagement was measured using an EEG metric: Engagement Index (EI). EI was calculated by taking the ratio of three power spectral density frequency bands (Freeman et al., 2004):

$$EI = \frac{Beta}{Alpha + Theta}$$

The rationale of EI is that increases in arousal and attention are reflected in the beta bandwidth while decreases are reflected in the alpha and theta bandwidths (Freeman, Mikulka, Prinzel, & Scerbo, 1999). It has been validated as a reliable measurement for engagement in adaptive system (Freeman et al., 2004; Mikulka, Scerbo, & Freeman, 2002). When first proposed, EI was calculated only using four sites: Cz, Pz, P3, and P4 (Pope, Bogart, & Bartolome, 1995), yet no solid rationality was given for excluding other channels. Later studies have used more or all channels available and obtained similar results (Chaouachi & Frasson, 2010; Freeman et al., 2000). Therefore, we have used all channels to calculate EI.

Performance metric was the same as described in RQ I. Raw TLX was used as a subjective metric for workload. In addition, two subscales of NASA-TLX: mental demand and effort were analyzed individually. Analyzing subscales was another common variation for NASA-TLX applications (Hart, 2006; Hoonakker et al., 2011). The consideration here is that pupil diameter and EI measurements might be associated with the two sub-dimensions respectively.

4.2.2 Data Analysis

4.2.2.1 Session variations

RQ II was interested in individual's performance changes between sessions. Therefore, all metrics were processed into session variations. The definitions of session variations are shown in Table 4.1. The first column gives the metric (and what it measures), 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 for performance was defined as:

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

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

Table 4.1 List of session variations

| Metrics | Symbol | Calculation |
|----------------------------------|-----------------|---|
| Performance metric | | |
| Performance | ΔP | $\Delta P_{i,j,k} = P_{i,j,k} - P_{i,j,k-1}$ |
| EEG metric | | |
| EI (engagement) | ΔE | $\Delta E_{i,j,k} = E_{i,j,k} - E_{i,j,k-1}$ |
| Eye tracking metric | | |
| Pupil diameter (mental workload) | ΔM | $\Delta M_{i,j,k} = M_{i,j,k} - M_{i,j,k-1}$ |
| Gaze entropy (scan strategy) | ΔS | $\Delta S_{i,j,k} = S_{i,j,k} - S_{i,j,k-1}$ |
| Subjective metrics | | |
| Raw TLX | $\Delta NASA$ | $\Delta NASA_{i,j,k} = NASA_{i,j,k} - NASA_{i,j,k-1}$ |
| NASA-TLX mental demand | $\Delta NASA^M$ | $\Delta NASA_{i,j,k}^M = NASA_{i,j,k}^M - NASA_{i,j,k-1}^M$ |
| NASA-TLX effort | $\Delta NASA^E$ | $\Delta NASA_{i,j,k}^E = NASA_{i,j,k}^E - NASA_{i,j,k-1}^E$ |

4.2.2.2 Improvement Labeling

In order to examine Q1 and Q3, the completion of exercise was categorized into improvement or non-improvement based on the value of ΔP . Instances where ΔP were below 0 were labeled as Decrease (non-improvement). For improvement, we assumed that ΔP was normally distributed and the number of Increase (improvement) should be the same as the number of Decrease in the other tail. Observations in the middle were small increases in performance and considered ambiguous, which were not be used for Q1 and Q3 but still valid in Q2.

4.2.2.3 Statistical Analysis

Three different analysis techniques were used to examine the 3 questions. Significance level for all analyses was set at $\alpha = 0.05$. When appropriate, p-values were corrected using Benjamini-Hochberg procedure.

Q1: The Analysis of Variance (ANOVA) was used to examine the difference of group means for two groups: Decrease and Increase. ANOVA is a widely used statistical technique for comparing group means (Montgomery, 2017). We seek to examine if the session variations of objective metrics were different for the two groups. Therefore, ΔM , ΔE and ΔS were used as response variables for ANOVA models respectively. Three subjective metrics: $\Delta NASA$, $\Delta NASA^M$ and $\Delta NASA^E$ were also analyzed using ANOVA.

Q2: The relationship between changes in performance and changes in objective metrics was examined using repeated measures correlation as described in RQ I. Correlations were calculated between ΔP and objective metrics/subjective metrics.

Q3: The same classification procedure as described in RQ I was used to explore the joint capability of objective metrics for predicting improvement. Specifically, the models used sessions variations of objective metrics to predict the label of improvement.

4.3 Results

4.3.1 Descriptive data

RQ II spanned about 3 months (including data collection in RQ I). There is one participant who only attended 1 session and was excluded from the analysis. Over the study period, a total of 26 sessions (294 exercises) were collected from 7 participants. 4 participants attended 3 sessions, 2 participants attended 5 sessions, and 1 participant attended 4 session. And all participants completed each exercise at least twice. Average values of all metrics from the first training attempt to the fifth are presented in Table 4.2. For most tasks, average performance became higher in later training attempts. Session variations were calculated for 212 exercises ($k \geq 2$). And among these 212 exercises, 61 observations have ΔP below 0 (which is the 30% quantile) and were labeled as

Decrease. Correspondingly, 61 observations with the highest ΔP (≥ 13 , which is the 70% quantile) were labeled as Increase (See Fig. 4.1).

Table 4.2 Mean value of all metrics across tasks and attempts

| Task | CT | | | | | PB | | | | | RR | | | | |
|---------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Attempt | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Performance metric | | | | | | | | | | | | | | | |
| Performance | 64.5 | 78.9 | 82.4 | 90.5 | 94.0 | 75.9 | 87.9 | 87.5 | 93.7 | 94.5 | 67.6 | 80.6 | 81.2 | 91.5 | 90.0 |
| Subjective metric | | | | | | | | | | | | | | | |
| NASA-TLX | 21.2 | 16.0 | 12.8 | 10.5 | 11.0 | 22.0 | 13.6 | 11.1 | 10.0 | 12.4 | 30.2 | 21.8 | 18.9 | 16.9 | 17.1 |
| EEG metric | | | | | | | | | | | | | | | |
| EI | 0.22 | 0.19 | 0.19 | 0.22 | 0.2 | 0.21 | 0.17 | 0.18 | 0.18 | 0.21 | 0.21 | 0.18 | 0.19 | 0.19 | 0.19 |
| Eye metrics | | | | | | | | | | | | | | | |
| Pupil diameter | 0.59 | 0.58 | 0.54 | 0.57 | 0.52 | 0.68 | 0.66 | 0.68 | 0.67 | 0.69 | 0.54 | 0.51 | 0.47 | 0.58 | 0.6 |
| Gaze entropy | 0.59 | 0.52 | 0.57 | 0.6 | 0.41 | 0.69 | 0.59 | 0.53 | 0.65 | 0.45 | 0.69 | 0.51 | 0.51 | 0.55 | 0.48 |

| Task | SS | | | | | DN | | | | | T | | | | |
|---------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Attempt | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| Performance metric | | | | | | | | | | | | | | | |
| Performance | 57.0 | 72.5 | 79.5 | 82.7 | 82.7 | 63.0 | 73.6 | 82.5 | 76.5 | 92.5 | 49.8 | 53.6 | 71.7 | 71.7 | 82.5 |
| Subjective metric | | | | | | | | | | | | | | | |
| NASA-TLX | 28.8 | 24.2 | 20.4 | 20.2 | 20.5 | 28.6 | 25.2 | 22.5 | 29.9 | 22.0 | 32.8 | 29.9 | 22.9 | 32.8 | 21.0 |
| EEG metric | | | | | | | | | | | | | | | |
| EI | 0.19 | 0.16 | 0.16 | 0.19 | 0.19 | 0.19 | 0.18 | 0.17 | 0.19 | 0.19 | 0.21 | 0.17 | 0.16 | 0.19 | 0.19 |
| Eye metrics | | | | | | | | | | | | | | | |
| Pupil diameter | 0.55 | 0.48 | 0.46 | 0.63 | 0.5 | 0.53 | 0.49 | 0.48 | 0.62 | 0.47 | 0.66 | 0.6 | 0.57 | 0.71 | 0.63 |
| Gaze entropy | 0.72 | 0.61 | 0.64 | 0.79 | 0.72 | 0.64 | 0.58 | 0.65 | 0.84 | 0.7 | 0.92 | 0.85 | 0.8 | 0.97 | 0.71 |

* CT: Camera Targeting, PB: Peg Board, RR: Ring and Rail, SS: Suture Sponge, DN: Dots and Needles, T: Tubes

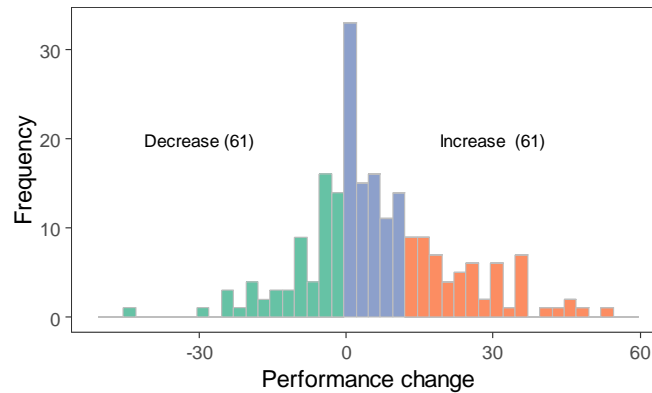


Fig. 4.1 Histogram of performance change

4.3.2 Objective Metrics and Improvement

The variations of objective metrics: ΔM (pupil diameter), ΔE (EI) and ΔS (gaze entropy) were hypothesized to be different under two conditions. ANOVA models were used to compare the mean between two groups: Decrease and Increase. F tests suggested there was significant difference between conditions for two sessions variations: ΔE ($F_{1,120} = 11.47, p < .001$) and ΔS ($F_{1,120} = 21.75, p < .001$). The mean and standard deviation of session variations under two conditions are shown in Table 4.3. In the Decrease group, the mean for ΔE and ΔS were both above zero while in the Increase group the mean were negative. Mean values of objective metrics variations by task are depicted in Fig. 4.2.

Table 4.3 Objective metrics variation and practice outcome

| Objective metrics | | ΔP : Decrease | | ΔP : Increase | |
|--------------------|----------------|-----------------------|-----|-----------------------|------|
| | | Mean | SD | Mean | SD |
| EEG metric | | | | | |
| ΔE | EI | .01 | .05 | -0.02 | 0.06 |
| Eye metrics | | | | | |
| ΔM | Pupil diameter | -.01 | .11 | -0.01 | 0.06 |
| ΔS | Gaze entropy | .05 | .17 | -0.1 | 0.18 |

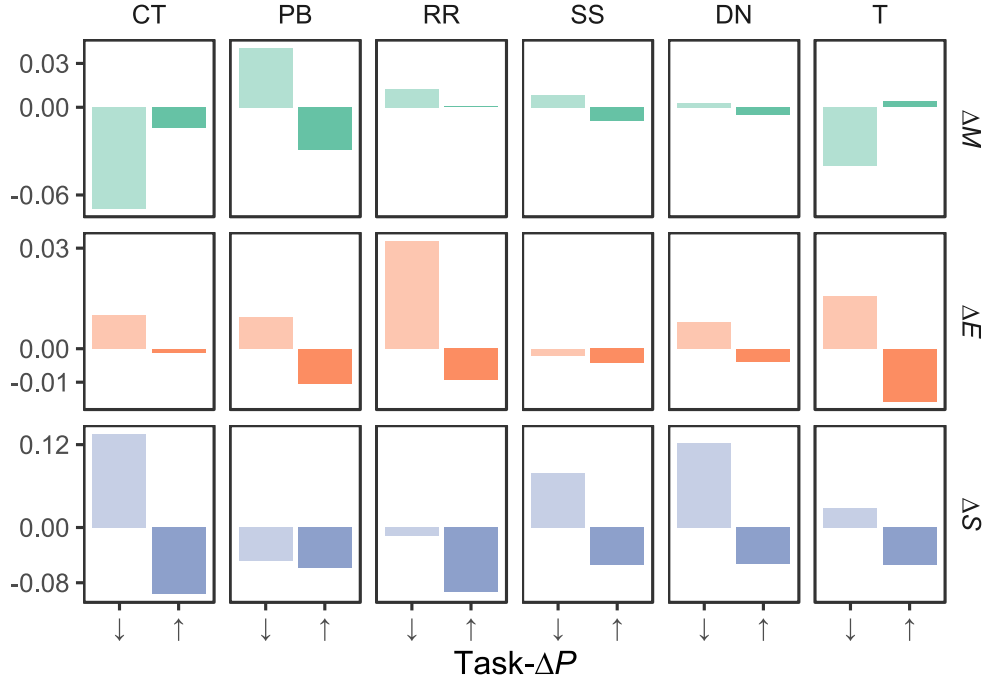


Fig. 4.2 Mean values of objective variations in two conditions by task

4.3.3 Objective Metrics and Performance Change

The repeated measures correlation test was used to explore whether objective metrics variations were correlated with the performance variation. And the results suggested that both ΔE ($r_{rm} = -.27, p < .001$) and ΔS ($r_{rm} = -.38, p < .001$) were significantly correlated with ΔP . The effect size was small for ΔE and medium for ΔS . The significant correlations indicated that larger changes in performance were likely to be accompanied by larger changes in EI and gaze entropy. Although the overall correlations were significant, there was variability among participants (Fig. 4.3).

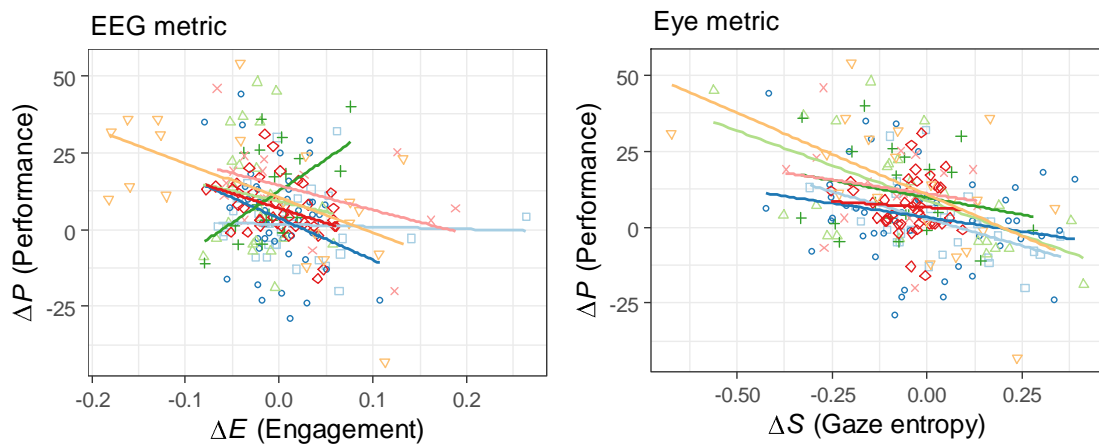


Fig. 4.3 Correlation between performance variation and objective metrics variation (colored and shaped by participants)

4.3.4 Classification of improvement

Logistic regression, Naïve Beys and SVM algorithms were used to train the models and predict the label of improvement. The models took 4 features as input: three metric session variations: ΔM , ΔE and ΔS ; and time interval between sessions. The 122 observations were partitioned into 3 sets with the size of 40, 40 and 42 for the 3-fold cross-validation. The average accuracy of validation was 68.5% and the average F1 score 0.67. The confusion matrix of the 3-fold cross validations for each algorithm are presented in

Table 4.4 (1)-(5). And the Naïve Bayes model has achieved the best performance with an average accuracy of 72.3% and average F1 score of 0.70.

Table 4.4 Improvement classification confusion matrix for testing dataset
(122 observations stratified into 3 folds)

(1) Logistic regression

| Logistic | | Actual class | | | |
|-----------------|----------|----------------|----------------|-------------|--------------|
| Predicted class | Decrease | Decrease | Increase | | |
| | | 33.0%±13.3% | 14.8%±2.8% | 67.8%±7.1% | |
| | Increase | True positive | False positive | Precision | |
| | | 17.0%±13.3% | 35.2%±2.8% | 70.8%±19.7% | |
| | | False negative | True negative | NPV | |
| | | 70.4%±5.6% | 70.4%±5.6% | 68.2%±12.5% | 0.66±0.17 F1 |
| | | Sensitivity | Specificity | Accuracy | score |

(2) Naïve Bayes

| Naïve Bayes | | Actual class | | | |
|-----------------|----------|----------------|----------------|-------------|-----------|
| Predicted class | Decrease | Decrease | Increase | | |
| | | 32.9%±9.4% | 10.6%±2.7% | 74.6%±9.4% | |
| | Increase | True positive | False positive | Precision | |
| | | 17.1%±9.4% | 39.4%±2.7% | 70.8%±13.1% | |
| | | False negative | True negative | NPV | |
| | | 65.9%±18.7% | 78.7%±5.5% | 72.3%±11.7% | 0.70±0.15 |
| | | Sensitivity | Specificity | Accuracy | F1 score |

(3) SVM (Linear Kernel)

| SVM-Linear | | Actual class | | | |
|-----------------|----------|----------------|----------------|-------------|--------------|
| Predicted class | Decrease | Decrease | Increase | | |
| | | 35.5%±12.4% | 19.8%±5.4% | 63.6%±4.9% | |
| | Increase | True positive | False positive | Precision | |
| | | 14.5%±12.4% | 30.2%±5.4% | 70.9%±15.1% | |
| | | False negative | True negative | NPV | |
| | | 60.5%±10.7% | 60.5%±10.7% | 65.7%±8.9% | 0.66±0.14 F1 |
| | | Sensitivity | Specificity | Accuracy | score |

(4) SVM (Gaussian Kernel)

| SVM-Gaussian | | Actual class | | | |
|-----------------|----------|----------------|----------------|-------------|--------------|
| Predicted class | Decrease | Decrease | Increase | | |
| | | 35.5%±12.4% | 17.3%±4.6% | 66.3%±7.8% | |
| | Increase | True positive | False positive | Precision | |
| | | 14.5%±12.4% | 32.7%±4.6% | 72.1%±16.0% | |
| | | False negative | True negative | NPV | |
| | | 65.5%±9.1% | 65.5%±9.1% | 68.2%±11.4% | 0.68±0.16 F1 |
| | | Sensitivity | Specificity | Accuracy | score |

Table 4.4 Continued

(5) SVM (Polynomial kernel)

| SVM-Poly | | Actual class | | | |
|-----------------|----------|----------------|----------------|------------|--------------|
| Predicted class | Decrease | Decrease | Increase | | |
| | | 33.0%±10.1% | 14.8%±6.8% | 69.4%±5.5% | |
| | Increase | True positive | False positive | Precision | |
| | | 17.0%±10.1% | 35.2%±6.8% | 69.0%±9.0% | |
| | | False negative | True negative | NPV | |
| | | 70.3%±13.6% | 70.3%±13.6% | 68.1%±6.6% | 0.66±0.12 F1 |
| | | Sensitivity | Specificity | Accuracy | score |

4.3.5 Subjective Metrics

Subjective metrics were also examined for Q1 and Q2. Similar to objective metrics, ANOVA and repeated measures correlations were used to test the relationships between ΔP and $\Delta NASA$, $\Delta NASA^M$, and $\Delta NASA^E$. ANOVA tests indicated that sessions variations were different for the two groups for all subjective metrics: $\Delta NASA$ ($F_{1,120} = 55.18$, $p < .001$), $\Delta NASA^M$ ($F_{1,120} = 11.60$, $p < .001$) and $\Delta NASA^E$ ($F_{1,120} = 36.51$, $p < .001$). And r_{rm} were also significant for all metrics, with larger, medium and small effect size respectively. Table 4.5 shows the r_{rm} , p-value and mean for the subjective metric variations. Mean values of subjective metric variations by task are presented in Fig. 4.4.

Table 4.5 Subjective metrics and improvement

| Subjective metrics | | r_{rm} | p-value | Mean | |
|--------------------|---------------|----------|---------|-----------------------|-----------------------|
| | | | | ΔP : Decrease | ΔP : Increase |
| $\Delta NASA$ | Raw TLX | -.51 | < .001 | 0.39 | -8.11 |
| $\Delta NASA^M$ | Mental Demand | -.27 | < .001 | -0.36 | -1.34 |
| $\Delta NASA^E$ | Effort | -.43 | < .001 | 0.05 | -1.53 |

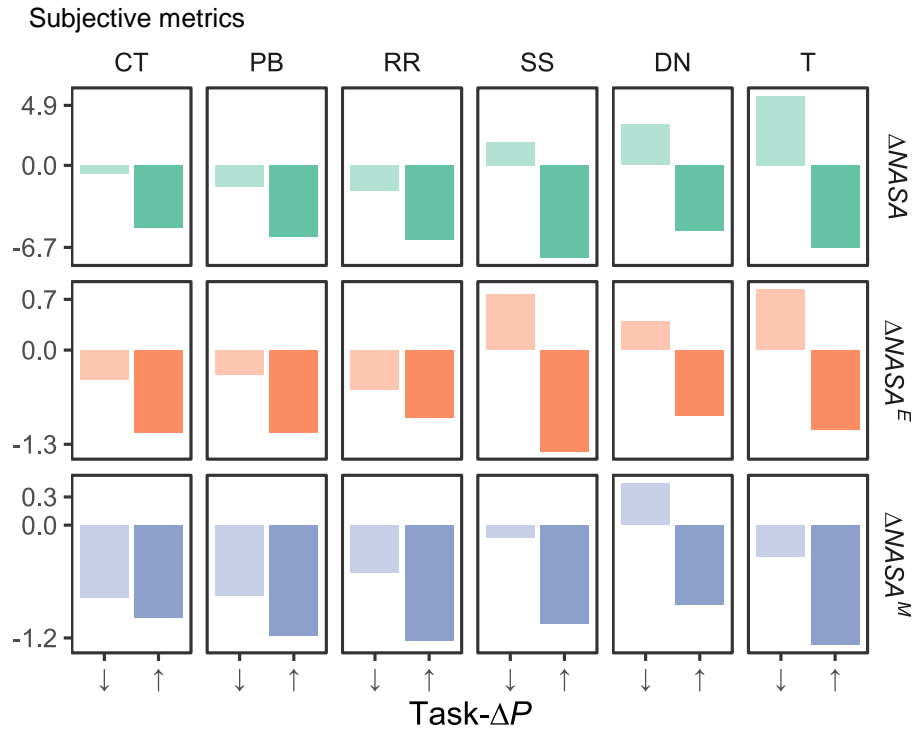


Fig. 4.4 Mean values of subjective variations in two conditions by task

4.4 Discussion

RQ II seeks to explain how performance changed during RAS training with objective metrics. Three metrics: pupil diameter, EI and gaze entropy were investigated considering changes in mental workload, engagement and gaze patterns. The first question examined was the difference in session variations between two conditions: improvement (Increase) and non-improvement (Decrease). The findings showed that sessions variations of EI and gaze entropy had different means in the two groups. The second question was whether changes in metrics are correlated with changes in task performance. Again, the results showed that sessions variations of EI and gaze entropy were correlated of variations of performance. The third question examined the feasibility of predicting improvement with session variations of objective metrics, and classification models have achieved a medium accuracy.

Gaze entropy showed a strong relationship with performance in that the mean variations (ΔS) was above 0 in the Increase group but below 0 in the Decrease group, and ΔS was also negatively

correlated with ΔP . In RQ I, gaze entropy has been proven to be a valid measurement of cognitive workload. Yet it essentially measured the gaze pattern (scan strategy) in RAS tasks, and gaze pattern could be impacted by high workload and task difficulty (See Discussion in RQ I). RQ II provided additional evidence that gaze pattern was related to performance in RAS tasks. Visual search is an indispensable step in surgery and studies on laparoscopic surgery have both qualitatively and quantitatively found that experts used a more effective gaze pattern than novice (Chetwood et al., 2012; Khan et al., 2012; Wilson et al., 2010). The task demands in this study are consistent with live robotic surgeries, where surgeons must rely on visual cues to complete the operation. These visual cues are delivered from the camera inside the patient that capture both current tissue states and robotic arms location. This information (e.g., current locations in respect to their desired target) is critical for planning actions necessary for completing the task goals. When searching for the target, trainees need to visually locate the target and also physically move controls to reach the target. Thus, eye tracking measures can directly provide data for understanding trainees' task performance and learning process.

The results for EI are similar to those for gaze entropy: ΔE was also negatively correlated with ΔP . The meaning of engagement varies as a function of study domain. In researches of vigilance task, task engagement was thought to be critical in sustaining attention and cognitive resources over time (Neigel, Dever, Claypoole, & Szalma, 2019) and a multivariate model factorized task engagement into energetic arousal, task motivation, and concentration (Matthews, Warm, & Smith, 2017). The engagement measured through EEG signals was more broadly described as reflection of information-gathering and allocation of attention (Stikic et al., 2014). And the EI metric adopted in this study reflected the state of alertness or attentiveness to task relevant stimuli (Freeman et al., 1999). Studies in vigilance task have stated that task engagement was positively correlated with task performance (Matthews et al., 2017). Our study focused on the how individual's engagement on the same task changed with practices. The results generally suggested that if the task performance was better than that in the last session, the engagement on the current session was likely to be lower. Based on the skill acquisition theory, there exists a series of sequenced stages from initial representation of knowledge to highly skilled behavior. Restructuring of procedural knowledge (knowing how to perform a process) will result in automatized knowledge (process performed correctly and rapidly) (DeKeyser, 2007). And the general agreement is that more

automatized knowledge is, the less attention the process requires and the less error-prone it is (Dekeyser & Criado, 2012). Therefore, the decreasing engagement can be interpreted as a reflection of knowledge atomization of tasks after training.

Pupil diameter did not show significant results for both Q1 and Q2. As discussed in RQ, pupil diameter was not correlated with workload measured by NASA-TLX. And there were two possible explanations. The first one is that various stimulus in RAS tasks have interfered with pupil dilation and the second one is that pupil diameter and NASA-TLX measured different aspects of workload. Apart from these explanations, another possibility is that there was a dissociation between performance and workload, with specific cognitive processes responsible for workload differing from those responsible for performance (Vidulich & Wickens, 1986; Yeh & Wickens, 1988).

However, by analyzing the subjective metrics we found that workload measured by NASA-TLX could be a significant explaining factor for performance, as examined by Q1 and Q2. The Raw TLX was a most common usage of NASA-TLX (Hart, 2006) and its session variations Δ NASA had the strongest correlation with ΔP . Nonetheless, trainees saw their performance scores before filling out NASA-TLX survey, which might be responsible for the correlation. We therefore seek to analyze some subscales, which is another feasible practice (Hart & Staveland, 1988; Hoonakker et al., 2011). Session variation in the Effort dimension had medium correlation with ΔP , resonating with results from EI. The correlation in Mental Demand dimension also had small correlation. It seemed that workload was responsible for changes in RAS performance, yet an alternative objective measurement is needed. Since mental workload for the same task could decrease after training, it will be beneficial to let surgeons practice procedures that may lead to high workload. Reducing workload in real-time is more complicated though. Recent study has proposed the possibility to provide dynamic gaze clues based on experts' data and provide guidance in surgery (Fichtel et al., 2019), but more studies are needed for a concrete and feasible intervention.

Finally, supervised machine learning techniques have achieved an average accuracy of 68.5% in predicting improvement and non-improvement. Compared with classification accuracy in RQ I, these results are not ideal. There is still a significant source for changes in performance that were

not explained by the current metrics, suggesting for unknown factors or factors difficult to assess through physiological and behavioral measures.

Task performance in RAS training is an important indicator of surgeon's mastery of techniques. Studies in various domains have suggested that objective measures are correlated with task performance by reflecting workload, vigilance and other latent factors. RQ II compared how much objective metrics have changed over training sessions and how much performance have changed. Findings provided evidence that objective metrics can explain changes in performance during RAS training. In the analysis, we did not consider where the participants were on their learning curve (i.e. how many sessions they have already attended). Since we are analyzing changes between sessions, previous experience was not supposed to impact the result. For example, when participants are very skilled and unlikely to achieve even higher performance, their objective metrics should also remain stable. All participants in this study were either medical students or residents, and the results might be different for experienced surgeons. For one things, experienced surgeons may already possess most basic skills and will achieve high performance in the beginning or very shortly (Dulan et al., 2012b). Yet, for surgeons experienced in laparoscopy, it is possible that they will perceive high workload when adapting to a new technique (Lee et al., 2014). In this study, we have observed that about 30% of the time participants would perform worse than last session. This result may indicate that RAS skills learned through simulated training can decline fast (Zhang & Sumer, 2013) and the current RAS training curriculum can be improved for more effective training. Objective metrics can be used to explain behaviors and cognitive states associated with improvement and help estimate room for improvement, and they may also provide measurements for live RAS procedure, where performance scores might not be available.

5. CONCLUSION AND FUTURE RESEARCH

RAS is a growing and promising part of surgery and it is envisaged that most surgery can and will be performed by robotic surgery in the future (Hashizume & Tsugawa, 2004). In 2016, installations of the most widespread robotic surgical system: da Vinci, have risen by 21% to >2500 units worldwide, and robotic procedures leaped by 25% to >450 000, in urology, gynecology and visceral surgery (Rassweiler et al., 2017). With increases in the usage, the technical advantages and clinical benefits are more widely acknowledged. However, the skills in RAS are unique and not derivative from either open or laparoscopic surgery, and the most practical and efficient of way of acquiring the basic skills is through simulation instead of the operating room (Bric, Lombard, Frelich, & Gould, 2016). Therefore, enhancing the assessments of RAS training through an HFE perspective can facilitate surgeon's mastery of RAS skills and improve patient safety.

Overall, the findings from this research emphasized the use of objective metrics to understand workload and performance in RAS. In RQ I, the use of eye tracking metrics would provide a fairly accurate estimation of both task workload and perceived workload, with gaze entropy showing strongest relationship with workload. And the classification model further confirmed the feasibility of using eye tracking signals for real-time workload level classification. The assessment of workload can augment RAS procedures from two perspectives. Firstly, the continuous workload assessment provides information about each step in a RAS procedure, so that the training procedure can be adjusted according to individual's needs. This assessment can be further applied in operating rooms to provide feedback for surgeons and advise them to avoid working while overloaded.

Findings from RQ II provided evidence that objective metrics can explain changes in performance during RAS training. The findings provide insights for physiological/behavioral changes accompanied by improvement. The engagement and scan strategy are the major factors that are related to changes in performance. Scan strategy plays an important role in surgical performance, and if the trainee's scan strategy improves over training, it indicates that the he or she is efficiently gaining RAS skills. Increasing engagement during training curriculum in contrast indicates that the trainee probably has not gained familiarity and dexterity. Still, the low classification accuracy

suggested that task performance is a complex result driven by multiple factors and some factors remain uncaptured. Objective metrics can be used to assess trainees' room for improvement, it may also provide measurements for live RAS procedure, where performance scores might not be available.

Despite the promising findings, a replication of this study should consider several issues. Due to the curriculum nature, task order in this study 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 use more advanced simulation tasks. In addition, the number of sessions and exercises for each participant was not specifically controlled in the study. Although neither RQ I and RQ II involved inter-participants comparison, having a consistent number of sessions and exercises for each individual can improve analysis accuracy. For RQ II, despite the extended data collection, the actual training duration for each task was still limited and even inadequate for observing asymptotic performance in the learning curve. Therefore, extra data collection can contribute to a more robust study. The da Vinci simulator used in this study is commercially available and used all over the world, 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 and explore the impact of task structure. And the eye tracking system could not synchronize eye gaze with the da Vinci screen like with a typical computer screen. Therefore, a dry lab study with more controls could consolidate the current findings.

Findings from this study opens some further question on physiological (and behavioral) metrics for future research. Gaze entropy, for example, was shown to be a sensitive measurement of workload in RQ I, but what this metric truly measures is the randomness of the scan pattern. Therefore, it is worth investigating whether the relationship between gaze entropy and workload depends on the specific visual demand of a task. Further study may want to explore interactions of visual skills, cognitive skills, and manual-manipulation skills required by different tasks on the objective metrics. Meanwhile, this study was performed in simulated training environment, and the transferability of the results to live surgery should be explored in the future.

APPENDIX A. RAS TASK ANALYSIS

Table A.1 Simulated robotic surgical tasks analysis

This is a high-level analysis which does not fully capture the magnitude of demands due to factors like movement/search distances and directions. a more rigorous method like Queueing Network-Model Human Processor (QN-MHP) will be needed for computational purpose.

| Camera Targeting (CT) | | | |
|-----------------------|-----------|--|--|
| Level | | 1 | 2 |
| Objective | | Focus the camera on different blue spheres spread across a broad pelvic cavity. | Maintain objective 1; pick up small cylinder under one sphere and transfer it to another sphere |
| Procedure | | 1. Search for sphere 2. Step on pedal to activate camera moving 3. Move both robotic arms to change camera view ^[1] 4. Grip both robotic claws to activate zooming 5. Move robotic arm to zoom into sphere Repeat 1-5 6 times | 1. Search for sphere 2. Step on pedal to activate camera moving 3. Move both robotic arms to change camera view 4. Grip both robotic claws to activate zooming 5. Move robotic arm to zoom into sphere 6. Release pedal and claws 7. Move arm to reach for cylinder 8. Grip one claw to pick up cylinder 9. Hold the claw 10. Search for next sphere 11. Step on pedal to activate camera moving 12. Move both robotic arms to change camera view 13. Grip both robotic claws to activate zooming 15. Move robotic arm to zoom into sphere 16. Release one claw to drop cylinder Repeat 1-16 4 times |
| Demand | Manual | Minimum 24 manual movements | Minimum 52 manual movements |
| | Visual | Minimum 6 exploratory visual search Minimum 12 fixations | Minimum 12 exploratory visual search Minimum 20 fixations |
| | Cognitive | Recognize signal/object Plan movement path | Recognize signal/object Plan movement path |
| Peg Board (PB) | | | |

| Level | | 1 | 2 |
|-----------|-----------|--|---|
| Objective | | Grasp rings on a vertical stand with the left hand and then pass them to the right hand before placing them on a peg | Same as objective 1 |
| Procedure | | <p>1. Find the ring that is flashing</p> <p><i>Optional (0-n times):</i> ^[2]</p> <p>2. Step on pedal to activate camera moving</p> <p>3. Move both robotic arms to change camera view</p> <p>4. Grip both robotic claws to activate zooming</p> <p>5. Move robotic arm to zoom</p> <p>6. Release pedal and claws</p> <p>7. Move arm to reach for ring</p> <p>8. Grip one claw to pick up ring</p> <p>9. Move one arm close to the other</p> <p>10. Release one claw</p> <p>11. Grip the other claw to transfer the ring</p> <p>12. Hold the claw</p> <p>13. Find the flashing peg</p> <p>14. Move arm to reach for the peg</p> <p>15. Release the claw to drop the ring</p> <p>Repeat 1-15 6 times</p> | <p>1. Searching the ring that is flashing</p> <p><i>Optional (1-n times):</i></p> <p>2. Step on pedal to activate camera moving</p> <p>3. Move both robotic arms to change camera view</p> <p>4. Grip both robotic claws to activate zooming</p> <p>5. Move robotic arm to zoom</p> <p>6. Release pedal and claws</p> <p>7. Move arm to reach for ring</p> <p>8. Grip one claw to pick up ring</p> <p>9. Move one arm close to the other</p> <p>10. Release one claw</p> <p>11. Grip the other claw to transfer the ring</p> <p>12. Hold the claw</p> <p>13. Search the flashing peg</p> <p><i>Optional (1-n times):</i></p> <p>14. Step on pedal to activate camera moving</p> <p>15. Move both robotic arms to change camera view</p> <p>16. Grip both robotic claws to activate zooming</p> <p>17. Move robotic arm to zoom</p> <p>18. Release pedal and claws</p> <p>19. Move arm to reach for the peg</p> <p>20. Release the claw to drop the ring</p> <p>Repeat 1-20 6 times</p> |
| Demand | Manual | Minimum 54 manual movements | Minimum 114 manual movements |
| | Visual | Minimum 12 exploratory visual search Minimum 18 fixations | Minimum 12 exploratory visual search Minimum 30 fixations |
| | Cognitive | Recognize signal/object Plan movement path | Recognize signal/object Plan movement path |

| Ring and Rail (RR) | | | | |
|--------------------|-----------|---|--|-------------------------------------|
| Level | | 1 | 2 | |
| Objective | | Move a ring along a twisted metal rod | Move 3 colored rings along 3 twisted and color-matched metal rod | |
| Procedure | | <div>1. Move arm to reach for the ring</div> <div>2. Grip one claw to pick up the ring</div> <div>3. Hold the claw</div> <div>4. Move arm to reach for the rod</div> <div><i>Optional (2-n times):</i></div> <div>5. Move arm to let ring go through the rod</div> <div>6. Release the claw to drop the ring at the end</div> | <div>1. Move arm to reach for ring</div> <div>2. Grip one claw to pick up ring</div> <div>3. Hold the claw</div> <div>4. Move arm to reach for the rod</div> <div>5. Move arm to let ring go through the rod</div> <div><i>Optional (2-n times):</i></div> <div>6. Step on pedal to activate camera moving</div> <div>7. Move both robotic arms to change camera view</div> <div>8. Grip both robotic claws to activate zooming</div> <div>9. Move robotic arm to zoom</div> <div>10. Release pedal and claws</div> <div><i>Optional (5-n times):</i></div> <div>11. Move arm to let ring go through the rod</div> <div>12. Release the claw to drop the ring at the end</div> <div>Repeat 1-12 3 times</div> | |
| Demand | Manual | Minimum 7 manual movements | Minimum 63 manual movements | |
| | Visual | Minimum 4 fixations | Minimum 12 fixations | |
| | Cognitive | Recognize signal/object Plan movement path | Recognize signal/object Plan movement path Plan the order of moving 3 rings Estimate the force and angle needed to move the ring without being impeded by rod | |
| Suture Sponge (SS) | | | | |
| Level | | 1 | 2 | 3 |
| Objective | | Drive needle through random targets on a deformable sponge | Same as objective 1 | Same as objective 1 |
| Procedure | | 1. Move arm to reach for the needle | 1. Move arm to reach for the needle | 1. Move arm to reach for the needle |

| | | | | |
|-----------------------|--|---|---|--|
| | | 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arm to reach for the target 6. Move arm to drive the needle into the target 7. Find the end-target 8. Move arm to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arm to reach the end of needle 11. Grip the claw to grip the needle 12. Move arm to pull the needle out of the sponge Repeat 1-12 10 times | 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arm to reach for the target 6. Move arm to drive the needle into the target 7. Find the end-target 8. Move arm to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arm to reach the end of needle 11. Grip the claw to grip the needle 12. Move arm to pull the needle out of the sponge <i>Optional (0-n times):</i> 13. Step on pedal to activate camera moving 14. Move both robotic arms to change camera view Repeat 1-14 8 times | 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arm to reach for the target 6. Move arm to drive the needle into the target 7. Find the end-target 8. Move arm to drive the needle puncture through the sponge and come out of the end-target 9. Release the claw 10. Move arm to reach the end of needle 11. Grip the claw to grip the needle 12. Move arm to pull the needle out of the sponge <i>Optional (0-n times):</i> 13. Step on pedal to activate camera moving 14. Move both robotic arms to change camera view Repeat 1-14 10 times |
| Demand | Manual | Minimum 120 manual movements | Minimum 96 manual movements | Minimum 120 manual movements |
| | Visual | Minimum 12 fixations | Minimum 12 fixations | Minimum 12 fixations |
| | Cognitive | Recognize signal/object Plan movement path Estimate the force and angle needed to drive the needle and hit the vertical end target | Recognize signal/object Plan movement path Estimate the force and angle needed to drive the needle and hit the vertical and diagonal end target | Recognize signal/object Plan movement path Estimate the force and angle needed to drive the needle and hit the vertical and diagonal end target |
| Dots and Needles (DN) | | | | |
| Level | 1 | | 2 | |
| Objective | Insert a needle through several pairs of targets that have various spatial positions | | Same as objective 1 | |
| Procedure | 1. Move arm to reach for the needle | | 1. Move arm to reach for the needle | |

| | | | |
|-----------|-----------|---|--|
| | | 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arm to reach for the target 6. Move arm to drive the needle into the target 7. Find the end-target 8. Move arm to drive the needle puncture through the pad and come out of the end-target 9. Release the claw 10. Move arm to reach the end of needle 11. Grip the claw to grip the needle 12. Move arm to pull the needle out of the sponge Repeat 1-12 7 times | 2. Grip one claw to pick up the needle 3. Hold the claw 4. Find the flashing begin-target 5. Move arm to reach for the target 6. Move arm to drive the needle into the target 7. Find the end-target 8. Move arm to drive the needle puncture through the end and come out of the end-target 9. Release the claw 10. Move arm to reach the end of needle 11. Grip the claw to grip the needle 12. Move arm to pull the needle out of the sponge <i>Optional (0-n times):</i> 13. Step on pedal to activate camera moving 14. Move both robotic arms to change camera view Repeat 1-14 6 times |
| Demand | Manual | Minimum 84 manual movements | Minimum 72 manual movements |
| | Visual | Minimum 21 fixations | Minimum 18 fixations |
| | Cognitive | Recognize signal/object Plan movement path Estimate the force and angle needed to drive the needle and hit the horizontal end target | Recognize signal/object Plan movement path Estimate the force and angle needed to drive the needle and hit the horizontal and diagonal end target |
| Tubes (T) | | | |
| Objective | | Drive needle through fixed targets on a cylindrical deformable structure | |
| Procedure | | 1. Search for the target 2. Move arm to reach for the cylinder 3. Grip claw to grip the edge of the cylinder 4. Hold the claw 5. Move arm to flip the cylinder 6. Find the flashing target 7. Move the other arm to reach for the needle 9. Grip the claw to pick up the needle 10. Hold the claw | |

| | | |
|--|-----------|--|
| | | 11. Move arm to drive the needle through the target 12. Release the claw 13. Move arm to reach the end of needle 14. Grip the claw to grip the needle 15. Move arm to pull the needle out of the target Repeat 1-15 8 times |
| Demand | Manual | Minimum 120 manual movements |
| | Visual | Minimum 8 exploratory visual search Minimum 24 fixations |
| | Cognitive | Recognize signal/object Plan movement path Plan the angle of holding the cylinder Estimate the force and angle needed to drive the needle and hit the target Estimate the force and angle needed to pull out the needle without hitting the cylinder |
| [1] Information about distance and movement direction is not included, which differs between tasks and levels [2] The minimum number of optional movements is based on optimal situation, which is rarely achieved in reality | | |

APPENDIX B. NASA-TLX SURVEY

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

| | | |
|------|------|------|
| Name | Task | Date |
|------|------|------|

Mental Demand How mentally demanding was the task?

Very Low
Very High

Physical Demand How physically demanding was the task?

Very Low
Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low
Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect
Failure

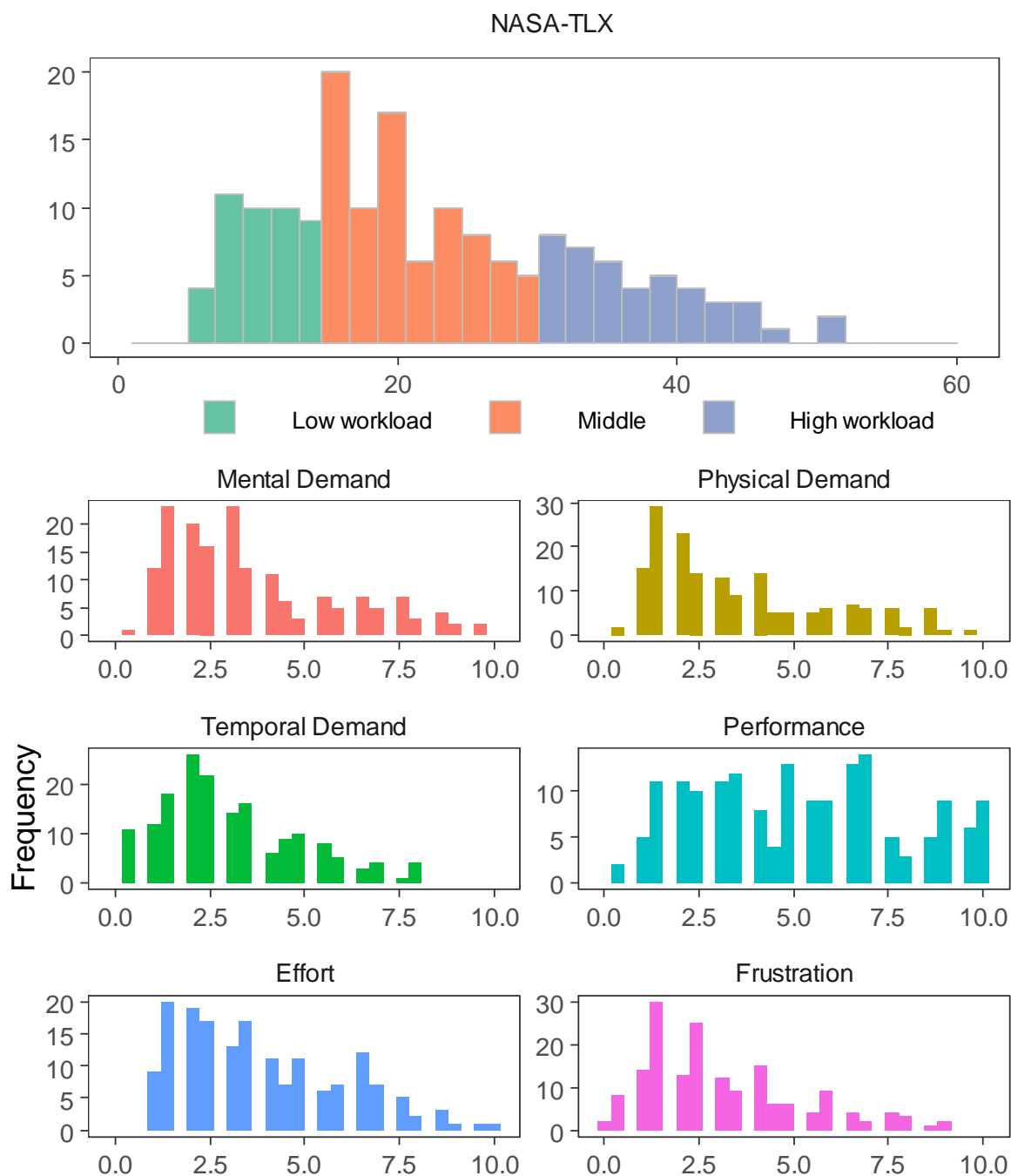
Effort How hard did you have to work to accomplish your level of performance?

Very Low
Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Very Low
Very High

APPENDIX C. NASA-TLX HISTOGRAMS



APPENDIX D. TABLE OF CORRELATIONS

Table A.2 Repeated correlation between eye metrics and NASA-TLX subscales

| NASA-TLX subscale | | Eye metrics | | | |
|---------------------|----------|----------------|--------------|-------------------|---------|
| | | Pupil diameter | Gaze entropy | Fixation Duration | PERCLOS |
| Mental Demand | r_{rm} | -.07 | .46 | .07 | .04 |
| | p | .360 | <.001 | .348 | .537 |
| Physical Demand | r_{rm} | -.13 | .40 | .09 | .03 |
| | p | .091 | <.001 | 0.266 | .683 |
| Temporal Demand | r_{rm} | -.19 | -.49 | -.01 | .08 |
| | p | .013 | <.001 | .923 | .302 |
| Overall Performance | r_{rm} | -.04 | .38 | .16 | .02 |
| | p | .573 | <.001 | .044 | .771 |
| Effort | r_{rm} | -.11 | .44 | .11 | -.01 |
| | p | .156 | <.001 | .158 | .887 |
| Frustration Level | r_{rm} | -.07 | .46 | .07 | .04 |
| | p | .352 | <.001 | .388 | .607 |

REFERENCES

- Abboudi, H., Khan, M. S., Aboumarzouk, O., Guru, K. A., Challacombe, B., Dasgupta, P., & Ahmed, K. (2013). Current status of validation for robotic surgery simulators – a systematic review. *BJU International*, *111*, 194–205.
- Abdelrahman, A. M., Bingener, J., Yu, D., Lowndes, B. R., Mohamed, A., McConico, A. L., & Hallbeck, M. S. (2016). Impact of single-incision laparoscopic cholecystectomy (SILC) versus conventional laparoscopic cholecystectomy (CLC) procedures on surgeon stress and workload: A randomized controlled trial. *Surgical Endoscopy*, *30*, 1205–1211.
- Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*, 735–739.
- Alzaharani, T., Haddad, R., Alkhayal, A., Delisle, J., Drudi, L., Gotlieb, W., ... Anidjar, M. (2013). Validation of the da Vinci Surgical Skill Simulator across three surgical disciplines: A pilot study. *Canadian Urological Association Journal*, *7*, E520–E529.
- Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review*, *22*, 425–438.
- Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., & Kneebone, R. (2010). The impact of stress on surgical performance: A systematic review of the literature. *Surgery*, *147*, 318–330.e6.
- Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*, 36–47.
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated Measures Correlation. *Frontiers in Psychology*, *8*, 456.
- Baldwin, C. L., & Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage*, *59*, 48–56.
- Ballantyne, G. H. (2002). The Pitfalls of Laparoscopic Surgery: Challenges for Robotics and Telerobotic Surgery. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, *12*, 1.
- Barbur, J. L., Harlow, A. J., & Sahraie, A. (1992). Pupillary responses to stimulus structure, colour and movement. *Ophthalmic and Physiological Optics*, *12*, 137–141.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276.
- Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*, 371–372.
- Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of Psychophysiology*, *2*.
- 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*, 289–300.
- Berguer, R., Smith, W. D., & Chung, Y. H. (2001). Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. *Surgical Endoscopy*, *15*, 1204–1207.

- Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007). *EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks*. 78, 14.
- Bokhari, M. B., Patel, C. B., Ramos-Valadez, D. I., Ragupathi, M., & Haas, E. M. (2011). Learning curve for robotic-assisted laparoscopic colorectal surgery. *Surgical Endoscopy*, 25, 855–860.
- Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44, 58–75.
- Bric, J. D., Lombard, D. C., Frelich, M. J., & Gould, J. C. (2016). Current state of virtual reality simulation in robotic surgery training: A review. *Surgical Endoscopy*, 30, 2169–2178.
- Brinkman, W. M., Luursema, J.-M., Kengen, B., Schout, B. M. A., Witjes, J. A., & Bekkers, R. L. (2013). Da Vinci Skills Simulator for Assessing Learning Curve and Criterion-based Training of Robotic Basic Skills. *Urology*, 81, 562–566.
- Cain, B. (2007). *A review of the mental workload literature*. Defence Research And Development Toronto (Canada).
- Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior Research Methods*, 41, 113–117.
- Card, S., Moran, T., & Newell, A. (1986). The model human processor- An engineering model of human performance. *Handbook of Perception and Human Performance*, 2.
- Carswell, C. M., Clarke, D., & Seales, W. B. (2005). Assessing mental workload during laparoscopic surgery. *Surgical Innovation*, 12, 80–90.
- Catchpole, K., Bisantz, A., Hallbeck, M. S., Weigl, M., Randell, R., Kossack, M., & Anger, J. T. (2018). Human factors in robotic assisted surgery: Lessons from studies ‘in the Wild.’ *Applied Ergonomics*. <https://doi.org/10.1016/j.apergo.2018.02.011>
- Chang, L., Satava, R. M., Pellegrini, C. A., & Sinanan, M. N. (2003). Robotic surgery: Identifying the learning curve through objective measurement of skill. *Surgical Endoscopy And Other Interventional Techniques*, 17, 1744–1748.
- Chaouachi, M., & Frasson, C. (2010). Exploring the Relationship between Learner EEG Mental Engagement and Affect. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (Vol. 6095, pp. 291–293). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Chen, L., Zhao, Y., Zhang, J., & Zou, J. (2015). Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Systems with Applications*, 42, 7344–7355.
- Chen, S., Epps, J., & Chen, F. (2013). Automatic and Continuous User Task Analysis via Eye Activity. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 57–66. New York, NY, USA: ACM.
- Chetwood, A. S., Kwok, K.-W., Sun, L.-W., Mylonas, G. P., Clark, J., Darzi, A., & Yang, G.-Z. (2012). Collaborative eye tracking: A potential training tool in laparoscopic surgery. *Surgical Endoscopy*, 26, 2003–2009.
- Cirett Galán, F., & Beal, C. R. (2012). EEG Estimates of Engagement and Cognitive Workload Predict Math Problem Solving Outcomes. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 51–62). Springer Berlin Heidelberg.
- Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, 16, 2349–2380.

- Colle, H. A., & Reid, G. B. (2005). Estimating a Mental Workload Redline in a Simulated Air-to-Ground Combat Mission. *The International Journal of Aviation Psychology*, 15, 303–319.
- Damos, D. (1991). *Multiple Task Performance*. CRC Press.
- Davis, W. T., Fletcher, S. A., & Guillamondegui, O. D. (2014). Musculoskeletal occupational injury among surgeons: Effects for patients, providers, and institutions. *Journal of Surgical Research*, 189, 207–212.e6.
- Dawson, A. P., McLennan, S. N., Schiller, S. D., Jull, G. A., Hodges, P. W., & Stewart, S. (2007). Interventions to prevent back pain and back injury in nurses: A systematic review. *Occupational and Environmental Medicine*, 64, 642–650.
- Debener, S., Minow, F., Emkes, R., Gandras, K., & Vos, M. de. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, 49, 1617–1621.
- 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.
- Di Stasi, Leandro L., Diaz-Piedra, C., Rieiro, H., Carrión, J. M. S., Berrido, M. M., Olivares, G., & Catena, A. (2016). Gaze entropy reflects surgical task load. *Surgical Endoscopy*, 30, 5034–5043.
- Di Stasi, Leandro L., Díaz-Piedra, C., Ruiz-Rabelo, J. F., Rieiro, H., Sanchez Carrion, J. M., & Catena, A. (2017). Quantifying the cognitive cost of laparo-endoscopic single-site surgeries: Gaze-based indices. *Applied Ergonomics*, 65, 168–174.
- Di Stasi, Leandro Luigi, Antolí, A., & Cañas, J. J. (2013). Evaluating mental workload while interacting with computer-generated artificial environments. *Entertainment Computing*, 4, 63–69.
- Diana, M., & Marescaux, J. (2015). Robotic surgery. *British Journal of Surgery*, 102, e15–e28.
- Dias, R. D., Ngo-Howard, M. C., Boskovski, M. T., Zenati, M. A., & Yule, S. J. (2018). Systematic review of measurement tools to assess surgeons' intraoperative cognitive workload. *BJS*, 105, 491–501.
- Dingemanse, N. J., & Dochtermann, N. A. (2013). Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, 82, 39–54.
- Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 103–130.
- Dulan, G., Rege, R. V., Hogg, D. C., Gilberg-Fisher, K. M., Arain, N. A., Tesfay, S. T., & Scott, D. J. (2012a). Developing a comprehensive, proficiency-based training program for robotic surgery. *Surgery*, 152, 477–488.
- Dulan, G., Rege, R. V., Hogg, D. C., Gilberg-Fisher, K. M., Arain, N. A., Tesfay, S. T., & Scott, D. J. (2012b). Proficiency-based training for robotic surgery: Construct validity, workload, and expert levels for nine inanimate exercises. *Surgical Endoscopy*, 26, 1516–1521.
- Egelund, N. (1982). Spectral analysis of heart rate variability as an indicator of driver fatigue. *Ergonomics*, 25, 663–672.
- ElBardissi, A. W., Wiegmann, D. A., Henrickson, S., Wadhera, R., & Sundt, T. M. (2008). Identifying methods to improve heart surgery: An operative approach and strategy for implementation on an organizational level. *European Journal of Cardio-Thoracic Surgery*, 34, 1027–1033.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.

- Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Rönning, J., Forlizzi, J. F., & Dey, A. K. (2014). Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. *2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB)*, 39–48.
- Feyen, R. G., & Liu, Y. (2001). The Queuing Network Model Human Processor (QNMHP): An Engineering Approach for Modeling Cognitive Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 45, 1733–1737.
- Fichtel, E., Lau, N., Park, J., Henrickson Parker, S., Ponnala, S., Fitzgibbons, S., & Safford, S. D. (2019). Eye tracking in surgical education: Gaze-based dynamic area of interest can discriminate adverse events and expertise. *Surgical Endoscopy*, 33, 2249–2256.
- Finnegan, K. T., Meraney, A. M., Staff, I., & Shichman, S. J. (2012). da Vinci Skills Simulator Construct Validation Study: Correlation of Prior Robotic Experience With Overall Score and Time Score Simulator Performance. *Urology*, 80, 330–336.
- Foot, J. R., & Valea, F. A. (2016). Robotic surgical training: Where are we? *Gynecologic Oncology*, 143, 179–183.
- Freeman, F. G., Mikulka, P. J., Prinzel, L. J., & Scerbo, M. W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology*, 50, 61–76.
- Freeman, F. G., Mikulka, P. J., Scerbo, M. W., Prinzel, L. J., & Clouatre, K. (2000). Evaluation of a Psychophysiologically Controlled Adaptive Automation System, Using Performance on a Tracking Task. *Applied Psychophysiology and Biofeedback*, 25, 103–115.
- Freeman, F. G., Mikulka, P. J., Scerbo, M. W., & Scott, L. (2004). An evaluation of an adaptive automation system using a cognitive vigilance task. *Biological Psychology*, 67, 283–297.
- Fuchs, K. H. (2002). Minimally Invasive Surgery. *Endoscopy*, 34, 154–159.
- Gabaude, C., Barakat, B., Jallais, C., Bonniaud, M., & Fort, A. (2012). Cognitive load measurement while driving. In: Human Factors: A view from an integrative perspective. In *Cognitive load measurement while driving. In: Human Factors: A view from an integrative perspective*. Human Factors and Ergonomics Society.
- Gawande, A. A., Thomas, E. J., Zinner, M. J., & Brennan, T. A. (1999). The incidence and nature of surgical adverse events in Colorado and Utah in 1992. *Surgery*, 126, 66–75.
- Gawande, A. A., Zinner, M. J., Studdert, D. M., & Brennan, T. A. (2003). Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*, 133, 614–621.
- Gilbreth, F. B., & Kent, R. T. (1911). *Motion study*. Constable London.
- Giulianotti, P. C., Coratti, A., Sbrana, F., Addeo, P., Bianco, F. M., Buchs, N. C., ... Benedetti, E. (2011). Robotic liver surgery: Results for 70 resections. *Surgery*, 149, 29–39.
- Greif, T. de, Lafeber, H., Oostendorp, H. van, & Lindenberg, J. (2009). Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation. *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience*, 219–228. Springer, Berlin, Heidelberg.
- Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., & Rao, R. P. N. (2008). Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 835–844. New York, NY, USA: ACM.
- Guru, K. A., Esfahani, E. T., Raza, S. J., Bhat, R., Wang, K., Hammond, Y., ... Chowriappa, A. J. (2015). Cognitive skills assessment during robot-assisted surgery: Separating the wheat from the chaff. *BJU International*, 115, 166–174.

- Guru, K. A., Shafiei, S. B., Khan, A., Hussein, A. A., Sharif, M., & Esfahani, E. T. (2015). Understanding Cognitive Performance During Robot-Assisted Surgery. *Urology*, 86, 751–757.
- Halverson, T., Estepp, J., Christensen, J., & Monnin, J. (2012). Classifying Workload with Eye Movements in a Complex Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56, 168–172.
- Harris Sr, R. L., Tole, J. R., Stephens, A. T., & Ephrath, A. R. (1981). *Visual scanning behavior and pilot workload*.
- Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50, 904–908. Sage Publications Sage CA: Los Angeles, CA.
- Hart, S. G., & Staveland, L. E. (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 *Springer Series in Statistics. The Elements of Statistical Learning* (pp. 193–224). Springer, New York, NY.
- He, X., Wang, L., Gao, X., & Chen, Y. (2012). The eye activity measurement of mental workload based on basic flight task. *IEEE 10th International Conference on Industrial Informatics*, 502–507.
- Hemal, A. K., Srinivas, M., & Charles, A. R. (2001). Ergonomic problems associated with laparoscopy. *Journal of Endourology*, 15, 499–503.
- Henelius, A., Hirvonen, K., Holm, A., Korpela, J., & Muller, K. (2009). Mental workload classification using heart rate metrics. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1836–1839.
- Henneman, E. A., Marquard, J. L., Fisher, D. L., & Gawlinski, A. (2017). Eye Tracking: A Novel Approach for Evaluating and Improving the Safety of Healthcare Processes in the Simulated Setting. *Simulation in Healthcare*, 12, 51–56.
- Henriksen, K., Dayton, E., Keyes, M. A., Carayon, P., & Hughes, R. (2008). *Understanding Adverse Events: A Human Factors Framework*. Agency for Healthcare Research and Quality (US).
- Hernandez, J. D., Bann, S. D., Munz, Y., Moorthy, K., Datta, V., Martin, S., ... Rockall, T. (2004). Qualitative and quantitative analysis of the learning curve of a simulated surgical task on the da Vinci system. *Surgical Endoscopy And Other Interventional Techniques*, 18, 372–378.
- Hignett, S. (2003). Intervention strategies to reduce musculoskeletal injuries associated with handling patients: A systematic review. *Occupational and Environmental Medicine*, 60, e6–e6.
- Hignett, Sue, Carayon, P., Buckle, P., & Catchpole, K. (2013). State of science: Human factors and ergonomics in healthcare. *Ergonomics*, 56, 1491–1503.
- Hoonakker, P., Carayon, P., Gurses, A. P., Brown, R., Khunlertkit, A., McGuire, K., & Walker, J. M. (2011). Measuring workload of ICU nurses with a questionnaire survey: The NASA Task Load Index (TLX). *IIE Transactions on Healthcare Systems Engineering*, 1, 131–143.

- Hubens, G., Ruppert, M., Balliu, L., & Vaneerdeweg, W. (2004). What Have we Learnt after Two Years Working with the Da Vinci Robot System in Digestive Surgery? *Acta Chirurgica Belgica*, 104, 609–614.
- Hung, A. J., Zehnder, P., Patil, M. B., Cai, J., Ng, C. K., Aron, M., ... Desai, M. M. (2011). Face, Content and Construct Validity of a Novel Robotic Surgery Simulator. *The Journal of Urology*, 186, 1019–1025.
- Intuitive Surgical, Inc. (2014). *Da Vinci S System User Manual* (User Manual No. 550516–06).
- 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, 1793–8201.
- Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, 34, 237–257.
- Jorna, P. G. A. M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, 36, 1043–1054.
- Jyothi, S., & Bhargavi, P. (2009). Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. 57. P.Bhargavi, S. Jyothi, 9.
- Kamzanova, A. T., Kustubayeva, A. M., & Matthews, G. (2014). Use of EEG workload indices for diagnostic monitoring of vigilance decrement. *Human Factors*, 56, 1136–1149.
- Kaul, S., Shah, N. L., & Menon, M. (2006). Learning curve using robotic surgery. *Current Urology Reports*, 7, 125–129.
- Kenney, P. A., Wszolek, M. F., Gould, J. J., Libertino, J. A., & Moinzadeh, A. (2009). Face, Content, and Construct Validity of dV-Trainer, a Novel Virtual Reality Simulator for Robotic Surgery. *Urology*, 73, 1288–1292.
- Khan, R. S. A., Tien, G., Atkins, M. S., Zheng, B., Panton, O. N. M., & Meneghetti, A. T. (2012). Analysis of eye gaze: Do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? *Surgical Endoscopy*, 26, 3536–3540.
- Khosrowabadi, R., Quek, C., Ang, K. K., Tung, S. W., & Heijnen, M. (2011). A Brain-Computer Interface for classifying EEG correlates of chronic mental stress. *The 2011 International Joint Conference on Neural Networks*, 757–762.
- Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011, September 13). A review of the use of psychophysiological methods in game research [Text]. https://doi.org/info:doi/10.1386/jgvw.3.3.181_1
- Klein, M. I., Warm, J. S., Riley, M. A., Matthews, G., Doarn, C., Donovan, J. F., & Gaitonde, K. (2012). Mental Workload and Stress Perceived by Novice Operators in the Laparoscopic and Robotic Minimally Invasive Surgical Interfaces. *Journal of Endourology*, 26, 1089–1094.
- Klein, M. I., Warm, J. S., Riley, M. A., Matthews, G., Gaitonde, K., & Donovan, J. F. (2008). Perceptual Distortions Produce Multidimensional Stress Profiles in Novice Users of an Endoscopic Surgery Simulator. *Human Factors*, 50, 291–300.
- Koca, D., Yıldız, S., Soyupek, F., Günyeli, İ., Erdemoglu, E., Soyupek, S., & Erdemoglu, E. (2015). Physical and Mental Workload in Single-Incision Laparoscopic Surgery and Conventional Laparoscopy. *Surgical Innovation*, 22, 294–302.
- Kohn, L. T., Corrigan, J., & Donaldson, M. S. (1999). *To err is human: Building a safer health system* (Vol. 6). National academy press Washington, DC.

- Lanfranco, A. R., Castellanos, A. E., Desai, J. P., & Meyers, W. C. (2004). Robotic Surgery. *Annals of Surgery*, 239, 14–21.
- Law, B., Atkins, M. S., Kirkpatrick, A. E., & Lomax, A. J. (2004). Eye Gaze Patterns Differentiate Novice and Experts in a Virtual Laparoscopic Surgery Training Environment. *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, 41–48. New York, NY, USA: ACM.
- Lean, Y., & Shan, F. (2012). Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 22, 177–187.
- Lee, D. S., Chong, T. W., & Lee, B. G. (2017). Stress Events Detection of Driver by Wearable Glove System. *IEEE Sensors Journal*, 17, 194–204.
- Lee, G. I., Lee, M. R., Clanton, T., Sutton, E., Park, A. E., & Marohn, M. R. (2014). Comparative assessment of physical and cognitive ergonomics associated with robotic and traditional laparoscopic surgeries. *Surgical Endoscopy*, 28, 456–465.
- Lee, J. Y., Mucksavage, P., Sundaram, C. P., & McDougall, E. M. (2011). Best Practices for Robotic Surgery Training and Credentialing. *Journal of Urology*, 185, 1191–1197.
- Lerner, M. A., Ayalew, M., Peine, W. J., & Sundaram, C. P. (2010). Does Training on a Virtual Reality Robotic Simulator Improve Performance on the da Vinci® Surgical System? *Journal of Endourology*, 24, 467–472.
- Lin, J. F., Frey, M., & Huang, J. Q. (2014). Learning curve analysis of the first 100 robotic-assisted laparoscopic hysterectomies performed by a single surgeon. *International Journal of Gynecology & Obstetrics*, 124, 88–91.
- Lowndes, B. R., & Hallbeck, M. S. (2014). Overview of Human Factors and Ergonomics in the OR, with an Emphasis on Minimally Invasive Surgeries. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 24, 308–317.
- Mack, M. J. (2001). Minimally Invasive and Robotic Surgery. *The Journal of the American Medical Association*, 285, 568–572.
- Marandi, R. Z., Samani, A., & Madeleine, P. (2018). The level of mental load during a functional task is reflected in oculometrics. In H. Eskola, O. Väisänen, J. Viik, & J. Hyttinen (Eds.), *EMBEC & NBC 2017* (pp. 57–60). Springer Singapore.
- Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of Eye-related Measures of Drivers' Mental Workload. *Procedia Manufacturing*, 3, 2854–2861.
- Marshall, S. P. (2000). *Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity*.
- Marucci, D. D., Shakeshaft, A. J., Cartmill, J. A., Cox, M. R., Adams, S. G., & Martin, C. J. (2000). Grasper trauma during laparoscopic cholecystectomy. *Australian and New Zealand Journal of Surgery*, 70, 578–581.
- Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors*, 57, 125–143.
- Matthews, G., Warm, J. S., & Smith, A. P. (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, 44–61.
- Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., Chera, B. S., & Marks, L. B. (2014). Relating physician's workload with errors during radiation therapy planning. *Practical Radiation Oncology*, 4, 71–75.

- Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., & Marks, L. B. (2013). Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. *Practical Radiation Oncology*, 3, e171–e177.
- Meshkati, N. (1988). Heart Rate Variability and Mental Workload Assessment. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (pp. 101–115). North-Holland.
- Meshkati, Najmedin, Hancock, P. A., Rahimi, M., & Dawes, S. M. (1995). *Techniques in mental workload assessment*.
- Metzenthin, P., Helfricht, S., Loerbroks, A., Terris, D. D., Haug, H. J., Subramanian, S. V., & Fischer, J. E. (2009). A one-item subjective work stress assessment tool is associated with cortisol secretion levels in critical care nurses. *Preventive Medicine*, 48, 462–466.
- Mikulka, P. J., Scerbo, M. W., & Freeman, F. G. (2002). Effects of a biocybernetic system on vigilance performance. *Human Factors*, 44, 654–664.
- Miller, S. (2001). Workload measures. *National Advanced Driving Simulator*. Iowa City, United States.
- Moglia, A., Ferrari, V., Morelli, L., Ferrari, M., Mosca, F., & Cuschieri, A. (2016). A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery. *European Urology*, 69, 1065–1080.
- Montgomery, D. C. (2017). *Design and analysis of experiments*. John Wiley & sons.
- Moore, L. J., Wilson, M. R., McGrath, J. S., Waine, E., Masters, R. S. W., & Vine, S. J. (2015). Surgeons' display reduced mental effort and workload while performing robotically assisted surgical tasks, when compared to conventional laparoscopy. *Surgical Endoscopy*, 29, 2553–2560.
- Moorthy, K., Munz, Y., Dosis, A., Bann, S., & Darzi, A. (2003). The effect of stress-inducing conditions on the performance of a laparoscopic task. *Surgical Endoscopy And Other Interventional Techniques*, 17, 1481–1484.
- Moorthy, K., Munz, Y., Dosis, A., Hernandez, J., Martin, S., Bello, F., ... Darzi, A. (2004). Dexterity enhancement with robotic surgery. *Surgical Endoscopy And Other Interventional Techniques*, 18, 790–795.
- Moorthy, Krishna, Munz, Y., Adams, S., Pandey, V., & Darzi, A. (2005). A Human Factors Analysis of Technical and Team Skills Among Surgical Trainees During Procedural Simulations in a Simulated Operating Theatre. *Annals of Surgery*, 242, 631–639.
- Morad, Y., Lemberg, H., Yofe, N., & Dagan, Y. (2000). Pupillography as an objective indicator of fatigue. *Current Eye Research*, 21, 535–542.
- Morris, R. K., Rayner, K., & Pollatsek, A. (1990a). Eye movement guidance in reading: The role of parafoveal letter and space information. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 268.
- Morris, R. K., Rayner, K., & Pollatsek, A. (1990b). Eye movement guidance in reading: The role of parafoveal letter and space information. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 268.
- Mosaly, P. R., Mazur, L. M., & Marks, L. B. (2017). Quantification of baseline pupillary response and task-evoked pupillary response during constant and incremental task load. *Ergonomics*, 60, 1369–1375.
- Moustafa, K., Luz, S., & Longo, L. (2017, June 4). *Assessment of Mental Workload: A Comparison of Machine Learning Methods and Subjective Assessment Techniques*. 30–50.

- Murai, K., Okazaki, T., Stone, L. C., & Hayashi, Y. (2007). A characteristic of a navigator's mental workload based on nasal Temperature. *2007 IEEE International Conference on Systems, Man and Cybernetics*, 3639–3643.
- Neigel, A. R., Dever, D. A., Claypoole, V. L., & Szalma, J. L. (2019). Task Engagement and the Vigilance Decrement Revisited: Expanding Upon the Work of Joel S. Warm Using a Semantic Vigilance Paradigm. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 61, 462–473.
- Nickel, P., & Nachreiner, F. (2003). Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload. *Human Factors*, 45, 575–590.
- Or, C. K., & Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, 7, 83–94.
- Palep, J. H. (2009). Robotic assisted minimally invasive surgery. *Journal of Minimal Access Surgery*, 5, 1–7.
- Palinko, O., Kun, A. L., Shyrovkov, 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. ACM.
- Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies*, 59, 185–198.
- 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, 2587–2593.
- Phé, V., Cattarino, S., Parra, J., Bitker, M.-O., Ambrogi, V., Vaessen, C., & Rouprêt, M. (2017). Outcomes of a virtual-reality simulator-training programme on basic surgical skills in robot-assisted laparoscopic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 13, e1740.
- Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40, 187–195.
- Putze, F., Jarvis, J., & Schultz, T. (2010). Multimodal Recognition of Cognitive Workload for Multitasking in the Car. *2010 20th International Conference on Pattern Recognition*, 3748–3751.
- Rajan, R., Selker, T., & Lane, I. (2016). Task Load Estimation and Mediation Using Psychophysiological Measures. *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 48–59. New York, NY, USA: ACM.
- Rassweiler, J. J., Autorino, R., Klein, J., Mottrie, A., Goezen, A. S., Stolzenburg, J.-U., ... Liatsikos, E. (2017). Future of robotic surgery in urology. *BJU International*, 120, 822–841.
- Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, 6, 31.
- Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology*, 11, 374–385.
- Richstone, R., & Richstone, J. (2010). Eye Metrics as an Objective Assessment of Surgical Skill. *Annals of Surgery*, 252, 177–182.
- Rieger, A., Stoll, R., Kreuzfeld, S., Behrens, K., & Weippert, M. (2014). Heart rate and heart rate variability as indirect markers of surgeons' intraoperative stress. *International Archives of Occupational and Environmental Health*, 87, 165–174.

- Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, 34, 259–287.
- Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, 53, 61–86.
- Sackier, J. M., & Wang, Y. (1994). Robotically assisted laparoscopic surgery. *Surgical Endoscopy*, 8, 63–66.
- Safren, M. A., & Chapanis, A. (1960). A critical incident study of hospital medication errors. *Nursing Research*, 9, 223.
- Sawilowsky, S. S. (2009). *New effect size rules of thumb*.
- Sawyer, D., Aziz, K. J., Backinger, C. L., Beers, E. T., Lowery, A., & Sykes, S. M. (1996). An introduction to human factors in medical devices. *US Department of Health and Human Services, Public Health Service, Food and Drug Administration, Center for Devices and Radiological Health*.
- Schreuder, H., Wolswijk, R., Zweemer, R., Schijven, M., & Verheijen, R. (2012). Training and learning robotic surgery, time for a more structured approach: A systematic review. *BJOG: An International Journal of Obstetrics & Gynaecology*, 119, 137–149.
- Schwalm, M., Keinath, A., & Zimmer, H. (2008). *Pupillometry as a method for measuring mental workload within a simulated driving task*.
- See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, 117, 230.
- Seixas-Mikelus, S. A., Kesavadas, T., Srimathveeravalli, G., Chandrasekhar, R., Wilding, G. E., & Guru, K. A. (2010). Face Validation of a Novel Robotic Surgical Simulator. *Urology*, 76, 357–360.
- Sexton, J., Thomas, E., & Helmreich, R. (2001). Error, Stress, and Teamwork in Medicine and Aviation: Cross Sectional Surveys. *Journal of Human Performance in Extreme Environments*, 6. <https://doi.org/10.7771/2327-2937.1019>
- Shah, P. R., Joseph, A., & Haray, P. N. (2005). Laparoscopic colorectal surgery: Learning curve and training implications. *Postgraduate Medical Journal*, 81, 537–540.
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5, 3–55.
- Sheikhzadeh, A., Gore, C., Zuckerman, J. D., & Nordin, M. (2009). Perioperating nurses and technicians' perceptions of ergonomic risk factors in the surgical environment. *Applied Ergonomics*, 40, 833–839.
- Singh, H., Bhatia, J. S., & Kaur, J. (2011). Eye tracking based driver fatigue monitoring and warning system. *Power Electronics (IICPE), 2010 India International Conference On*, 1–6. IEEE.
- So, W. K. Y., Wong, S. W. H., Mak, J. N., & Chan, R. H. M. (2017). An evaluation of mental workload with frontal EEG. *PLOS ONE*, 12, e0174949.
- Solovey, E. T., Zec, M., Garcia Perez, E. A., 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. ACM.
- Sommer, D., & Golz, M. (2010). Evaluation of PERCLOS based current fatigue monitoring technologies. *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 4456–4459. IEEE.

- Steinberg, P. L., Merguerian, P. A., Bihrlé, W., & Seigne, J. D. (2008). The Cost of Learning Robotic-Assisted Prostatectomy. *Urology*, 72, 1068–1072.
- Stemberger, J., Allison, R. S., & Schnell, T. (2010). Thermal Imaging as a Way to Classify Cognitive Workload. *2010 Canadian Conference on Computer and Robot Vision*, 231–238.
- Stikic, M., Berka, C., Levendowski, D. J., Rubio, R. F., Tan, V., Korszen, S., ... Wurzer, D. (2014). Modeling temporal sequences of cognitive state changes based on a combination of EEG-engagement, EEG-workload, and heart rate metrics. *Frontiers in Neuroscience*, 8. <https://doi.org/10.3389/fnins.2014.00342>
- Supe, A. N., Kulkarni, G. V., & Supe, P. A. (2010). Ergonomics in laparoscopic surgery. *Journal of Minimal Access Surgery*, 6, 31–36.
- Talamini, M. A., Chapman, S., Horgan, S., & Melvin, W. S. (2003). A prospective analysis of 211 robotic-assisted surgical procedures. *Surgical Endoscopy And Other Interventional Techniques*, 17, 1521–1524.
- Tanaka, A., Graddy, C., Simpson, K., Perez, M., Truong, M., & Smith, R. (2016). Robotic surgery simulation validity and usability comparative analysis. *Surgical Endoscopy*, 30, 3720–3729.
- Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39, 740–748.
- Tekkis, P. P., Senagore, A. J., Delaney, C. P., & Fazio, V. W. (2005). Evaluation of the Learning Curve in Laparoscopic Colorectal Surgery. *Annals of Surgery*, 242, 83–91.
- Tian, Y., Zhang, S., Wang, C., Yan, Q., & Chen, S. (2019). Eye Tracking for Assessment of Mental Workload and Evaluation of RVD Interface. In S. Long & B. S. Dhillon (Eds.), *Man-Machine-Environment System Engineering* (pp. 11–17). Springer Singapore.
- Tien, T., Pucher, P. H., Sodergren, M. H., Sriskandarajah, K., Yang, G.-Z., & Darzi, A. (2014). Eye tracking for skills assessment and training: A systematic review. *Journal of Surgical Research*, 191, 169–178.
- Tole, J. R. S. (1983). *Visual scanning behavior and pilot workload*. Retrieved from <https://ntrs.nasa.gov/search.jsp?R=19830025266>
- van Det, M. J., Meijerink, W. J. H. J., Hoff, C., Totté, E. R., & Pierie, J. P. E. N. (2009). Optimal ergonomics for laparoscopic surgery in minimally invasive surgery suites: A review and guidelines. *Surgical Endoscopy*, 23, 1279–1285.
- Veelen, M. A. van, Jakimowicz, J. J., & Kazemier, G. (2004). Improved physical ergonomics of laparoscopic surgery. *Minimally Invasive Therapy & Allied Technologies*, 13, 161–166.
- Verhage, R. J., Hazebroek, E. J., Boone, J., & Van, H. R. (2009). Minimally invasive surgery compared to open procedures in esophagectomy for cancer: A systematic review of the literature. *Minerva Chirurgica*, 64, 135.
- Vidulich, M. A., & Wickens, C. D. (1986). Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments. *Applied Ergonomics*, 17, 291–296.
- Wahr, J. A., Prager Richard L., Martinez Elizabeth A., Salas Eduardo, Seifert Patricia C., Groom Robert C., ... Nussmeier Nancy A. (2013). Patient Safety in the Cardiac Operating Room: Human Factors and Teamwork. *Circulation*, 128, 1139–1169.
- Walter, C., Schmidt, S., Rosenstiel, W., Gerjets, P., & Bogdan, M. (2013). Using Cross-Task Classification for Classifying Workload Levels in Complex Learning Tasks. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 876–881.

- Wang, X., & Xu, C. (2016). Driver drowsiness detection based on non-intrusive metrics considering individual specifics. *Accident Analysis & Prevention*, 95, 350–357.
- Wetzel, C. M., Kneebone, R. L., Woloshynowych, M., Nestel, D., Moorthy, K., Kidd, J., & Darzi, A. (2006). The effects of stress on surgical performance. *The American Journal of Surgery*, 191, 5–10.
- Whittaker, G., Aydin, A., Raison, N., Kum, F., Challacombe, B., Khan, M. S., ... Ahmed, K. (2015). Validation of the RobotiX Mentor Robotic Surgery Simulator. *Journal of Endourology*, 30, 338–346.
- Wierwille, W. W., Wreggit, S. S., Kirn, C. L., Ellsworth, L. A., & Fairbanks, R. J. (1994). *Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. Final report.*
- Wilson, G. F., & Russell, C. A. (2003). Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. *Human Factors*, 45, 635–644.
- Wilson, M. R., McGrath, J., Vine, S., Brewer, J., Defriend, D., & Masters, R. (2010). Psychomotor control in a virtual laparoscopic surgery training environment: Gaze control parameters differentiate novices from experts. *Surgical Endoscopy*, 24, 2458–2464.
- Wilson, M. R., Vine, S. J., Bright, E., Masters, R. S. W., Defriend, D., & McGrath, J. S. (2011). Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study. *Surgical Endoscopy*, 25, 3731–3739.
- Wottawa, C. R., Genovese, B., Nowroozi, B. N., Hart, S. D., Bisley, J. W., Grundfest, W. S., & Dutson, E. P. (2016). Evaluating Tactile Feedback in Robotic Surgery for Potential Clinical Application using an Animal Model. *Surgical Endoscopy*, 30, 3198–3209.
- Xie, A., & Carayon, P. (2015). A systematic review of human factors and ergonomics (HFE)-based healthcare system redesign for quality of care and patient safety. *Ergonomics*, 58, 33–49.
- Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 30, 111–120.
- Yokoi, H., Chen, J., Desai, M. M., & Hung, A. J. (2018). Impact of Virtual Reality Simulator in Training of Robotic Surgery. In Ashok K. Hemal & M. Menon (Eds.), *Robotics in Genitourinary Surgery* (pp. 183–202). Cham: Springer International Publishing.
- Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, 58, 1–17.
- Yu, D., Dural, C., Morrow, M. M. B., Yang, L., Collins, J. W., Hallbeck, S., ... Forsman, M. (2017). Intraoperative workload in robotic surgery assessed by wearable motion tracking sensors and questionnaires. *Surgical Endoscopy*, 31, 877–886.
- Yu, D., Lowndes, B., Morrow, M., Kaufman, K., Bingener, J., & Hallbeck, S. (2016). Impact of novel shift handle laparoscopic tool on wrist ergonomics and task performance. *Surgical Endoscopy*, 30, 3480–3490.
- 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, 1565–1574.
- Zhang, N., & Sumer, B. D. (2013). Transoral Robotic Surgery: Simulation-Based Standardized Training. *JAMA Otolaryngology–Head & Neck Surgery*, 139, 1111–1117.

- Zheng, B., Cassera, M. A., Martinec, D. V., Spaun, G. O., & Swanström, L. L. (2010). Measuring mental workload during the performance of advanced laparoscopic tasks. *Surgical Endoscopy*, 24, 45.
- Zheng, B., Jiang, X., & Atkins, M. S. (2015). Detection of Changes in Surgical Difficulty: Evidence From Pupil Responses. *Surgical Innovation*, 22, 629–635.
- Zheng, B., Jiang, X., Tien, G., Meneghetti, A., Panton, O. N. M., & Atkins, M. S. (2012). Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surgical Endoscopy*, 26, 2746–2750.
- Zhou, J., Jung, J. Y., & Chen, F. (2015). Dynamic Workload Adjustments in Human-Machine Systems Based on GSR Features. In J. Abascal, S. Barbosa, M. Fetter, T. Gross, P. Palanque, & M. Winckler (Eds.), *Human-Computer Interaction – INTERACT 2015* (pp. 550–558). Springer International Publishing.
- Zihni, A. M., Ohu, I., Cavallo, J. A., Cho, S., & Awad, M. M. (2014). Ergonomic analysis of robot-assisted and traditional laparoscopic procedures. *Surgical Endoscopy*, 28, 3379–3384.