Artificial Intelligence Tutoring Compared with Expert Instruction in Surgical Simulation Training. - A Randomized Controlled Trial

Ali M. Fazlollahi, BSc

Department of Experimental Surgery

Faculty of Medicine and Health Sciences

McGill University

Montréal, Canada

August 2021

A thesis submitted to McGill University in partial fulfilment of the requirement of the degree of Master of Science.

TABLE OF CONTENTS

Background

The Study

Thesis Summary

Tables

Figures

Appendix

ABSTRACT

Background

Resection of subpial tumors adjacent to critical brain areas is one of the most high-stakes surgical procedures that lacks opportunities for safe deliberate practice. Furthermore, performance-based assessment in neurosurgical apprenticeship is inefficient and vulnerable to subjectivity. Utilizing Artificial Intelligence (AI) to classify neurosurgical psychomotor expertise, we developed the first AI-powered tutor in neurosurgical simulation training, known as the Virtual Operative Assistant (VOA), to augment technical skills acquisition training by providing intelligent feedback.

Objectives

Determine how learning with the VOA compares with training by remote expert instruction in performing virtual reality brain tumor resections and experiencing emotions and cognitive load.

Methods

A multi-institutional randomized controlled trial compared VOA's automated audiovisual metric-based feedback vs remote verbal debriefing with expert instruction and no-feedback controls. Medical students performed six simulated subpial brain tumor resections: five practice attempts followed with feedback and one complex realistic attempt evaluated skill retention and transfer. A deep learning model, Intelligent Continuous Expertise Monitoring System (ICEMS Expertise Score) and blinded Objective Structured Assessment of Technical Skills (OSATS) evaluated performance. Participants reported emotions before, during and after training and completed a cognitive load questionnaire following training.

Results

Seventy medical students from four institutions were randomly assigned to VOA (n=23), Instructor (n=24), and Control (n=23) Groups. 350 practice attempts were assessed by ICEMS, and 70 realistic attempts were evaluated by ICEMS and OSATS. During practice, VOA training resulted in a significant improvement of participants’ Expertise Scores that was on average 0.66 (95% CI 0.55-0.77) and 0.65 (95% CI 0.54-0.77) points higher than Instructor and Control Groups (p<.001). Realistic attempt’s average Expertise Score was significantly higher in the VOA group compared to instructor and control groups (mean difference 0.53; 0.49, respectively, p<.001). VOA and instructor group’s realistic attempt OSATS ratings were not significantly different. OSATS ratings demonstrated that VOA feedback resulted in significantly higher overall performance, economy of movement, and respect for tissue compared to control while expert instruction significantly improved instrument handling, respect for tissue, and economy of movement compared to control. There was no significant between-groups difference in cognitive load, positive-, and negative-activating emotions.

Conclusion

VOA’s quantitative automated benchmark feedback demonstrated superior performance outcome, improved skill transfer, with equivalent OSATS ratings and similar cognitive and affective responses compared to remote expert instruction.

RÉSUMÉ

Contexte

La résection de tumeurs subpiales adjacentes à des zones cérébrales critiques est l'une des procédures chirurgicales aux enjeux les plus élevés, qui manque de possibilités de pratique délibérée sûre. De plus, l'évaluation basée sur la performance dans l'apprentissage de la neurochirurgie est inefficace et vulnérable à la subjectivité. En utilisant l'intelligence artificielle (IA) pour classifier l'expertise psychomotrice neurochirurgicale, nous avons développé le premier tuteur alimenté par l'IA dans la formation neurochirurgicale par simulation, connu sous le nom d'Assistant Opératoire Virtuel (VOA), pour augmenter la formation à l'acquisition de compétences techniques en fournissant un retour intelligent.

Objectifs

Déterminer comment l'apprentissage avec le VOA se compare à la formation par l'instruction d'un expert à distance en effectuant des résections de tumeurs cérébrales en réalité virtuelle et en ressentant des émotions et une charge cognitive.

Méthodes

Un essai contrôlé randomisé multi-institutionnel a comparé le retour d'information automatisé basé sur des mesures audiovisuelles du VOA à un débriefing verbal à distance avec instruction d'un expert et à des contrôles sans retour d'information. Des étudiants en médecine ont effectué six résections simulées de tumeurs cérébrales subpiales : cinq tentatives d'entraînement suivies d'un retour d'information et une tentative réaliste complexe ont permis d'évaluer la rétention et le transfert des compétences. Un modèle d'apprentissage profond, le système intelligent de surveillance continue de l'expertise (ICEMS Expertise Score) et l'évaluation objective structurée des compétences techniques (OSATS) en aveugle ont évalué les performances. Les participants ont fait part de leurs émotions avant, pendant et après la formation et ont rempli un questionnaire sur la charge cognitive après la formation.

Résultats

Soixante-dix étudiants en médecine de quatre institutions ont été répartis au hasard dans les groupes VOA (n=23), Instructeur (n=24) et Contrôle (n=23). 350 tentatives de pratique ont été évaluées par ICEMS, et 70 tentatives réalistes ont été évaluées par ICEMS et OSATS. Au cours de la pratique, la formation VOA a entraîné une amélioration significative des scores d'expertise des participants qui étaient en moyenne 0,66 (IC 95 % 0,55-0,77) et 0,65 (IC 95 % 0,54-0,77) points plus élevés que ceux des groupes instructeur et témoin (p<.001). Le score moyen d'expertise de la tentative réaliste était significativement plus élevé dans le groupe VOA par rapport aux groupes instructeur et contrôle (différence moyenne 0,53 ; 0,49, respectivement, p<.001). Les notes OSATS de la tentative réaliste du groupe VOA et du groupe instructeur n'étaient pas significativement différentes. Les évaluations OSATS ont démontré que le feedback de la VOA a entraîné une performance globale, une économie de mouvement et un respect des tissus significativement plus élevés par rapport au groupe témoin, tandis que l'instruction experte a amélioré de manière significative la manipulation des instruments, le respect des tissus et l'économie de mouvement par rapport au groupe témoin. Il n'y avait pas de différence significative entre les groupes en ce qui concerne la charge cognitive, les émotions positives et négatives.

Conclusion

Le feedback quantitatif automatisé de VOA a démontré un résultat de performance supérieur, un transfert de compétences amélioré, avec des évaluations OSATS équivalentes et des réponses cognitives et affectives similaires par rapport à l'enseignement expert à distance.

DEDICATION AND PREFACE

For all those who are currently battling with, or carrying the scars of, brain tumors and other neurological diseases.

This thesis is original work by the candidate, and it is structured in a manuscript-based format.

The abstract of the study, titled “Artificial Intelligence Tutoring Versus Expert Instruction in Surgical Simulation Training: Randomized Controlled Trial” has been accepted for an oral presentation in the Research and Innovation Session at the Simulation Summit of the Royal College of Physicians and Surgeons of Canada for November 2021.

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my supervisor and mentor, Dr. Rolando Del Maestro, who trusted me with carrying out this thesis project in his lab. You inspired me to pursue excellence in medicine and research and taught me that the most important thing we leave behind are "the dreams that our names inspire, and the works that make our names a symbol for admiration." Thank you for your confidence in my skills and the opportunities you provided me over the past couple of years. I look forward to continuing to learn from you in my future academic endeavours and medical career. I am thankful to my co-supervisor, Dr. Jason Harley, who taught me how to bridge theory and practice in research and build collaborations in academia. I would also like to acknowledge members of my research advisory committee, Dr. Hadil Al-Jallad, Dr. Livia Garzia, Dr. Reza Forghani, and Dr. Roy Dudley for their insightful feedback and support during my training. I admire your dedication to science and medicine and feel privileged for sharing my thesis project with you.

Conducting a randomized controlled trial during the COVID-19 pandemic was made possible by a collective effort and thanks to our team of multidisciplinary experts. I feel honoured for the opportunity to lead this project and learn from your talents. To my colleagues and collaborators, Dr. Recai Yilmaz, Dr. Mohamad Bakhaidar, Dr. Ahmad Alsayegh, and Dr. Alexander Winkler-Schwartz, I am grateful for your expertise and hard work that enabled our team to accomplish this work and I thank you for providing me with ample opportunities to learn about neurosurgery. I would like to thank all members of the Neurosurgical Simulation and Artificial Intelligence Learning Centre for sharing their multidisciplinary experience and helping me accrue knowledge and exposure in multiple domains from engineering to deep learning. I'm also thankful to the active participating audience of the Neurosim Group's Friday Forum

meetings, whose brilliant thoughts continue to inspire me. Thank you for listening to my presentations and providing feedback on how to improve my talks. Notably, I would like to thank all medical students who participated in this study and the individuals who worked tirelessly at the Montreal Neurological Institute's Security and Facilities services to ensure that this trial was conducted safely during this pandemic. You've shown incredible commitment to your profession and the advancement of science during this difficult time. It was my absolute pleasure to meet each one of you.

Finally, I would like to acknowledge my family and friends for their unwavering support during this difficult time. You pushed me to challenge myself and invest my efforts in activities that were beyond my comfort zone. To my parents, who as immigrants sacrificed so much to provide an enriching upbringing that exposed me to underserved populations around the world and inspired me to strive for the best and bring hope to those who need it most. You are my role-models and what I have achieved so far has only been possible because of your ongoing love and support. Thank you for showing me a glimpse of this world and encouraging me to pursue my dreams.

AUTHOR CONTRIBUTIONS

The candidate led this trial, contributed to all aspects of the study, including the design of the trial protocol, instructor training, data acquisition, statistical analysis, results interpretation, and manuscript writing. The candidate had full access to all the data and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Dr. Recai Yilmaz and Mr. Ian Langleben assisted in running the trial and providing technical knowledge. Dr. Yilmaz provided insight on statistical analysis and supported the assessment of performance data utilizing the LSTM model.

Dr. Mohamad Bakhaidar and Dr. Ahmad Alsayegh contributed to the instruction of participants and performing the blinded OSATS assessments.

Dr. Winkler-Schwartz provided important neurosurgical insight on the design of the instructor group including the feedback content for the VOA and the instructor's feedback script.

Mr. Nykan Mirchi assisted in optimizing the VOA feedback prior to recruitment.

Ms. Nicole Ledwos provided statistical insight.

Dr. Bajunaid and Dr. Sabbagh contributed to the development of components of the realistic tumor resection scenario.

Dr. Jason Harley provided expert insight on the design of questionnaires, analysis, and interpretation of emotion results, and provided high-level feedback on educational theories.

Dr. Rolando Del Maestro supervised the project plan and was primarily in charge of the study's overall direction. He acquired the ethics approval, provided the neurosurgical expertise and resources necessary to, design, conduct, and interpret the results of the study. He assisted in recruitment and provided ongoing support at every step of the trial.

ABBREVIATIONS

3D	Three Dimensional
AI	Artificial Intelligence
ANN	Artificial Neural Network
CBME	Competency-Based Medical Education
COVID-19	Coronavirus disease 2019
EPA	Entrustable Professional Activity
ICEMS	Intelligent Continuous Expertise Monitoring System
KNN	K-Nearest Neighbour
LSTM	Long Short-Term Memory
OSATS	Objective Structured Assessment of Technical Skills
O-SCORE	Ottawa Surgical Competency Operating Room Evaluation
PEARLS	Promoting Excellence And Reflective Learning in Simulation
PGY	Post-Graduate Year
SRL	Self-Regulated Learning
SVM	Support Vector Machine
VOA	Virtual Operative Assistant

INTRODUCTION

In competency-based medical education, the use of simulation technologies has become a central tool to augment patient safety, enable repetition of challenging tasks, and provide opportunities for deliberate practice.[1] Neurosurgery is a branch of medicine known for its high stakes and complex procedures, where small errors can result in significant patient morbidity.[2] For example, an incorrectly performed subpial resection of a brain tumor near the motor cortex can result in healthy tissue damage leading to a patient's loss of movement. This is a commonly performed procedure that requires competency in multiple areas including bimanual psychomotor proficiency and the appreciation of important anatomical structures.[3] In fact, technical performance has been demonstrated to account for up to one quarter of the variability in patients' post-operative complications and is directly associated with patient outcomes.[4,5] Therefore, it is critical to provide learners with sufficient training that ensures their technical mastery.

Employing virtual reality on high fidelity surgical simulators, researchers and educators can replicate the visual, auditory, and haptic experience of challenging procedures for practice without posing imminent risk to patients. Furthermore, computers running these simulated tasks can record large multimodal datasets on the user's interaction with the virtual world, such as the position of instruments and the volumes of blood emitted, or tumor and brain resected.[6] Utilizing artificial intelligence (AI) methods to analyze and compare datasets from experts and novices, our team demonstrated an ability to accurately classify individuals based on established performance metrics.[7-11] These metrics are measurable components of performance that act as features for prediction or classification by machine learning algorithms, and by relying on transparent AI and expert consultation, our team has ensured that these metrics are understandable and teachable.[12]

Besides the advantage of streamlining objective assessment of technical competency, researchers can use machine learning algorithms to generate intelligent tutoring systems. An intelligent tutoring system refers to an autonomous pedagogical agent that evaluates students' progress and provides personalized formative feedback to enhance performance.[13] In medical education, there have been a number of intelligent tutoring systems developed for teaching conceptual knowledge[14,15], however, the difficulty in measuring performance posed a challenge to developing systems for teaching procedural skills. With advances in virtual reality simulation, recording and evaluating performance by intelligent machines has become possible. As such, our team developed the first intelligent tutoring system for teaching technical skills in surgical simulation.[16]

The Virtual Operative Assistant (VOA) utilizes a type of a machine learning classifier algorithm known as a linear support vector machine, to predict learners' competency in safety and instrument movement, and automatically provides metric-specific audiovisual feedback to enhance their proficiency where lacking.[16] Although it promises to provide learners with an efficient platform for deliberate practice with formative feedback, it is unclear if learning with the system is effective and leads to enhanced performance. The system's ability in fostering an appropriate cognitive and affective learning environment also remains to be elucidated. The aim of this thesis is to investigate VOA's effectiveness as a learning platform and compare it with conventional instruction from an expert. The findings reported here can inform residency programs' decisions in supplementing their curriculum with intelligent tutors.

BACKGROUND

Simulation in Surgical Education

According to a 2018 report commissioned by the Canadian Patient Safety Institute, over the next 30 years, medical accidents will take the lives of roughly 24000 Canadians in acute care and cost our health care system 1.3 billion dollars every year.[17] From these incidents, surgical errors leave the costliest burden, and accidents in brain surgery specifically are most likely to leave a long-lasting dent in a patient's quality of life. Simulation has proven to be useful in high-risk professions, such as aviation and the military, for preparing pilots and special operation officers to train for and anticipate complex situations before they arise.[18-20] The use of simulation in surgery has not kept pace with its development in other fields, but presently, it is a field of much intensive research.[21]

Simulation enables the replication of real-life scenarios, procedures, and operations which can occur in the workplace. It provides a risk-free environment where learners can explore, take actions, and importantly make mistakes without any real consequences. Although there are various degrees of fidelity in simulation, from simple benchtop models to highly immersive virtual reality, the aim is to place learners in familiar environments while requiring them to make decisions in relation to the scenario's circumstances. Because of its highly standardized, reproducible, and competency-specific features, simulation-based training has become an area of great promise for assessment and feedback in surgical training.[22-24]

Simulation-based training is empirically rooted in several educational theories. Foremost, it offers potentially unlimited opportunities for repetition, the backbone of deliberate practice.[25] The popularized 10,000-hour rule predicates that individuals can master any task if they spend roughly that amount of time practicing their craft.[26] With the current 80-hour per week resident work-hour restrictions, mandated by the American Council of Graduate Medical Education to improve residents' fatigue and burnout,[27,28] not only are residents less exposed to operative procedures[29] but it would take them 125 weeks of uninterrupted practice to master a technical procedure. Furthermore, the 10,000-hour rule can be misleading because repetition alone is insufficient to lead to mastery. Vince Lombardi's famous quote that "only perfect practice makes perfect" highlights focused repetition with expert feedback for attaining expert performance.[30,31]

In self-regulated learning (SRL), the path to mastery is illuminated by students' growing awareness of their performance objectives and ability to set specific, measurable, attainable, relevant, and time-based goals.[32] Because simulation tasks often have competency-specific learning objectives, learners are more likely to develop the metacognitive abilities, such as forethought and self-reflective observation, to set specific goals and critically evaluate their performance towards those competencies.[33,34] Finally, the realism of simulation training promotes situated learning where learners consolidate their knowledge through immersion and active participation in the professional context in which the theory is applied.[35] In-situ simulation incorporates the sociocultural elements of situated learning and enables multidisciplinary teams to learn together in communities of practice.[36,37]

In designing highly immersive simulation training, it is essential to consider the intervention's emotional and cognitive demands to ensure effective acquisition of skills. Learning has a significant affective component that is often under-recognized,[38] and simulation has been shown to elicit various emotional states, including stress.[39] This is important because achievement emotions have been demonstrated to influence learners' academic and performance outcomes.[40,41] Students' affective states can initiate situationally appropriating information processing and self-regulating strategies that in turn focus attention, motivation, and cognitive resources on achievement related activities.[42,43] As a result, strength of emotions can shape learners' behaviour in how they approach goal-oriented tasks.

In general, emotions are categorized based on their valence (subjective feeling) and arousal (level of physiological activation).[44] Positive activating emotions, such as hope and enjoyment, are positively related to achievement, while negative deactivating emotions, such as boredom and hopelessness, generally impair achievement.[45] Positive deactivating emotions, such as relaxation and relief, and negative activating emotions, such as anger and anxiety, can result in both adaptive and maladaptive behavioural responses that support learning depending on their strength and the context of learning.[46] For example, feeling slightly anxious before an exam may motivate students to invest more time in preparing, whereas feeling overly anxious may lead to panic and an inability to perform.

Even though learning involves emotions, it is predominantly an intellectual task. According to the information processing theory, learning is a cognitive process that leads to a permanent change of knowledge or skill.[47] According to this theory, we learn by processing novel information in our short-term memory and incorporating them with previously established schemas stored in our long-term memory. Schemas are domain-specific knowledge constructs that require cognitive resources to build.[47] Multimodal input from our sensory organs first enter our working memory that selects relevant information to create and organize representative mental models. These models are then integrated with existing schemas and get encoded in our long-term memory.

Cognitive theory of learning introduces three assumptions. First is the notion of dual coding, which suggests that humans have two separate channels for processing visual and auditory information. Second assumption is that we have a limited capacity in the amount of information that can be processed in each channel at one time.[48] Third, is the notion of active processing where engaging in active learning is done by selectively attending to relevant incoming information, thus by virtue, ignoring irrelevant information.[49] Integrating these cognitive assumptions with information processing theory leads to the basis of cognitive load.[50]

According to this theory, we have limited cognitive resources thus learning may be impaired when the task demands exceed available cognitive capacity.[51,52] This theory introduces three types of cognitive load: Germane, intrinsic, and extrinsic load. Germaine load is the generative cognitive processing needed for making sense of the information, and it is impacted by the learner's motivation to learn and their strategies to organize and integrate the information. Intrinsic load is the inherent demands of learning the material. A topic with a higher intrinsic cognitive load refers to one with complex and more interconnected components. Extrinsic load is the extraneous cognitive demand imposed on the learners that requires them to separate relevant from non-relevant information. It is related to instruction design, how the information is presented, and has no bearing on the intrinsic difficulty of the concepts.

Although the inherent difficulty of a subject and a learner's motivation are beyond the control of instructors, there are some strategies to reduce extrinsic load through effective instructional design.[50-53] Removing seductive details, such as background music, reduces extraneous load according to the coherence principle, while segmentation (i.e., chunking related information into meaningful parts) has a positive effect on learning and transfer.[54] Highly immersive simulation training involves many interactive information elements which is reported to disproportionately increase multimodal processing and task complexity for novice learners.[55] Because this can influence learners' performance and learning outcomes,[56] it is important to consider cognitive load in the design and evaluation of novel pedagogical platforms.

Simulation in Neurosurgery

Surgical simulators can be divided into two general categories: Physical or virtual simulations. Physical simulations allow for direct manipulation of either real biologic tissue or synthetic objects (e.g., 3D printed models, box trainers). Biologic models range from human cadavers, used for example in spinal decompression simulation training,[57] to ex-vivo animal tissues, such as chicken wings for practicing microvascular anastomosis[58] or bovine brain to practice brain tumor resection.[59] These models provide excellent realism of the anatomical structures and the haptic experience of tissue manipulation, but they are costly to preserve, non-reusable, and may raise ethical, health, and safety issues. Synthetic objects, such as a gel-based 3D printed Willis' Circle for aneurysm training,[60] are less costly to produce and maintain, are mass-producible, and thus more accessible, but they often lack the dynamic realism of living tissue. This is especially important for neurosurgery where interpreting tactile and kinesthetic information with surgical instruments is essential, because the small operating field may leave insufficient visual feedback for navigation or tissue discrimination.[61,62]

Virtual simulations involve interaction with computer-generated visual, auditory, or tactile information that can either be supplemented to the real physical environment, as in the case of augmented reality, or entirely encompass the environment like in virtual reality.[63] Augmented reality provides learners with relevant information, such as live imaging,[64] or annotations, such as the accurate insertion path for a pedicle screw,[65] that can be used in remote instruction.[66] However, limitations such as fatigue and cognitive overload are important considerations for their successful integration in surgical practice. Although virtual reality simulators are less accessible due to their high initial cost,[67] they offer several advantages. First, they are highly immersive, so learners are engaged in a specific surgical scenario that is designed to create the audiovisual and haptic feeling of that operation. Second, the virtual reality simulations are consistent and repeatable that makes them both a useful tool for potentially unlimited practice opportunities and for conducting randomized controlled trials. Finally, because virtual reality simulators are computer-based systems, large volumes of data are recorded on users' interaction in the virtual environment that contain valuable information on the trainee's performance.[68-70]

The NeuroVR (CAE Healthcare, Montreal, Canada), formerly known as NeuroTouch (Figure 1), is a high-fidelity virtual reality simulator with haptic feedback that recreates the audiovisual and haptic experience of various neurosurgical procedures.[6,71] On this platform, learners look through a simulated neurosurgical microscope and interact with the virtual environment utilizing both hands and simulated instruments. The simulation scenarios illustrate the realistic anatomy, authentic dynamic movements (e.g., bleeding), and accurate physical properties (e.g., density) of the biologic tissues they represent.[72] As such, the face and content validity of this platform – i.e., its ability to represent the real-life surgical procedure and require the specific bimanual psychomotor skills for that procedure, respectively – has been reported by expert consultants. Furthermore, utilizing machine learning methods to evaluate the collected data from the simulation trial, its construct validity – i.e., the simulator's ability to differentiate expertise groups – has been established.[7,71]

Among other neurosurgical and spinal procedures, this platform offers two subpial tumor resection scenarios; a basic "practice" scenario (Figure 2) that covers the fundamental competencies and a "realistic" scenario (Figure 3) that is more complex and can be used to evaluate skill retention.[72,73] Both simulations involve the removal of a gioma-like primary human brain tumor using a simulated aspirator in the dominant hand and simulated bipolar forceps in the non-dominant hand. Previous piloted studies have shown that most experts and novices can complete the practice scenario in 5 minutes and the realistic scenario in 13 minutes.[7]

With the new training demands caused by the coronavirus disease 2019 (COVID-19) pandemic,[74] some authors have highlighted the role of simulation in neurosurgery to help these challenges.[75,76] Diminished case volume and increased learner time spent away from the operating room due to social distancing measures have led to innovative remote-based solutions through simulation to prevent further disruption in training.[77-79] Although there are numerous distance simulation opportunities for practicing clinical knowledge, such as online virtual patients,[80] practicing hands-on technical skills at home is more challenging. Some residency programs provided simulation kits that include physical models and box trainers for residents to practice at home. Junior surgical residents at Stanford University, for example, received silicone-based 3D printed hematoma models with the appropriate surgical instruments. Learners practiced this procedure at home, followed an online module that included the case description, and had the opportunity to discuss with a senior resident or consultant upon completion.[79]

There is evidence to suggest that distributed self-regulated training at home using the Fundamentals of Laparoscopy box trainer leads to an equivalent performance outcome compared to centralized instructor-regulated training for residents.[81] Early learners such as medical students, however, are more likely to require supportive instructions at early stages of training that ensure successful "scaffolding" – i.e., provide temporary support until a learner can carry out the task.[82] As such, using simulation methods for training junior learners, providing expert feedback and instruction through either augmented reality headsets or videotelephony software is an appropriate remote-based methodology.[66,83] Providing remote instruction requires an objective method to assess performance and trained instructors who can reliably use this tool.

Performance Assessment in Surgery – The Convention

Throughout the history of western medicine, surgery has been regarded as a craft and the notion of technical skill, despite being a characteristic element of a master surgeon, has taken different meanings based on distinct cultural and historical contexts. As such, similar to other crafts, the standards of excellence have changed through time.[84]

New technology and innovation have played critical roles in defining an expert surgical performance. For example, before the advent of general anaesthesia in mid-19th century, the speed of an operation and the surgeon's agility marked a masterful surgeon who could minimize patients' suffering during painful procedures. Robert Liston, a 19th century Scottish surgeon became a prominent figure of this era for performing limb amputations in just under a few minutes. However, with William Morton's demonstration of diethyl ether's anaesthetic effects in 1846, the audience in the surgical theatre observed a new form of performance: movements that occurred at a slower tempo, focused on precision for procedures that became more invasive.[84]

As a craft, learning surgery followed the apprenticeship paradigm, where trainees observed and learned from a skilled master. Like other skill-dependent vocations, the Halstedian model of "see one, do one, teach one" allowed learners to gradually assume more responsibility and require less supervision until they became capable of teaching the skill themselves.[85] However, in the modern world this model shows inefficiencies that researchers, program directors, and national licensing bodies have begun to address by developing competency-based medical education (CBME) curriculums.[86] In Canada, for example, residents' progress in the professional setting is measured through successful accomplishment of several Entrustable Professional Activities (EPAs). An EPA is a key task, specific to a specialty or a procedure, that a trainee can be trusted to perform in the health care context once sufficient competence has been demonstrated.[87] These activities attempt to bridge the apprenticeship model with the CBME framework by providing a structure for on-the-job assessment of learners' progression and facilitating feedback related to steps associated with a specific clinical task.[88] The EPAs help to standardize the core competencies required for a certain surgical procedure and ensure consistency in surgeons' procedural expertise.[87]

However, for the training of basic technical skills, utilizing EPAs alone presents challenges. Foremost, procedural definitions of specific EPAs provide little information on composites of technical quality. They also remain prone to bias and subjectivity, require expert consultant time for supervision and assessment, and impend on the availability of patient cases, that result in lower reproducibility, higher cost and may expose patients to increased risk.

As a result, criteria-based assessment methods, such as Global Rating Scales (GRS), the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE), or the Objective Structured Assessment of Technical Skills (OSATS), help to overcome these challenges and reduce the subjectivity in technical assessment of EPAs.[89,90] By defining specific operative qualities to observe in the learner's performance and providing a Likert scale for rating, these methods increase agreement among evaluators on what constitutes an expert surgical performance and help in directing feedback towards actions that lead to a better operative quality.[23]

From the criteria-based assessment methods, the OSATS visual rating has emerged as a versatile tool in technical skill assessment for various surgical subspecialities and has demonstrated predictive validity by establishing technical skills as an independent predictor of patient outcomes.[4,91] As one of the first tools developed to evaluate surgical skills objectively, it has been subject to numerous studies in multiple surgical subspecialties, and as a result, it is often regarded as one of the gold-standard tools for technical assessment in surgery.[92] OSATS's versatility lies in defining technical competency based on general operative qualities. The original scale includes seven domains of performance that are rated visually on a 5-point Likert scale. These include: 1) Respect of tissue, 2) Time and motion, 3) Instrument handling, 4) Knowledge of instruments, 5) Flow of operation, 6) Use of assistants, and 7) Knowledge of specific procedure.[89] These categories comprise multiple aspects of a surgical performance and are not restricted to a single procedure.[92]

In surgical training, the OSATS can be used for both summative and formative assessment. Summative assessment involves high-stakes decisions regarding an individual's progress, such as certification or the completion of residency.[23] For example, post-graduate year 1 (PGY1) general surgery residents at multiple institutions, including the University of Toronto, are required by their programs to complete the Fundamentals of Laparoscopic Surgery examination that includes a multi-station technical skill summative assessment based on the OSATS.[93] Formative assessment involves low-stakes evaluations aimed at improving trainee learning in various domains through appropriate feedback. Because routine use of criteria-based assessment in formative evaluation remains limited to instructor availability, it can be labour-intensive and costly to implement by training programs.[23] As such, the utility of OSATS in composing formative feedback for learners and its resulting effect on subsequent performance is understudied.[94]

In neurosurgery, the OSATS global rating scale was demonstrated to be a feasible and reliable tool of performance assessment in the operating room. In one study, 299 procedures by 8 pediatric neurosurgery residents were evaluated intraoperatively by one of the 6 supervising attending surgeons. This group demonstrated that faculty members scored senior residents

(PGY6) significantly higher than junior residents (PGY3-4) for the "knowledge of instruments" domain. When procedure difficulty was considered, for expert-level procedures, such as pineal tumor resection or revascularization surgery for Moyamoya, senior residents' scores from attending surgeons were significantly higher than the scores given to junior residents.[95] Furthermore, evidence from our group demonstrated that a modified OSATS visual rating scale can be reliably used to assess virtual reality simulations of neurosurgical procedures and that some performance domains of the OSATS demonstrate significant correlations with some of the simulator's performance metrics.[96]

As a result, in this project, our group utilized the OSATS visual rating scale to aid instructors of the study in assessing on-screen performances of medical students during virtual reality brain tumor operations. The scale was tested for internal consistency and raters were trained prior to recruitment to enhance inter-rater reliability. Because the simulated operations were performed without an assistant and participants were not evaluated on their knowledge, three OSATS domains, "Knowledge of instruments", "Use of assistants", and "Knowledge of specific procedure", were eliminated for being inapplicable to this study. Instead, two categories, namely "Overall" and "Hemostasis", were introduced to assess the overall operative quality and participants' ability to successfully control bleeding, respectively.[96] Both categories were previously studied and found to be relevant for simulated neurosurgical procedures. Furthermore, because of a time limit to perform the operations, "Time and motion" was adjusted to "Economy of movements".

Performance Assessment in Surgery – The New

Artificial intelligence (AI) is a broad term which refers to computers that can learn about the world flexibly, make inferences about what they see and hear, and achieve human-like understanding of information. Machine learning is a subset of AI that refers to the process where computers learn to recognize patterns in data from the examples provided to them and improve their performance using classification algorithms.[97] Machine learning algorithms can make predictive models that take in large amounts of information and classify the output into different categories. A k-nearest neighbour (KNN), artificial neural networks (ANN), and support vector machines (SVM) are typical machine learning algorithm structures, and each contain unique characteristics that gives them advantage for certain types of data.[9] Deep learning is a type of machine learning that consists of an ANN with multiple hidden layers. This allows algorithms to analyze more complex relationships within data and even identify novel features that enhance their decision making.[97]

AI has proven to be a useful tool in enabling humans to understand hidden complexities and make accurate predictions to keep up with our expanding ability to generate, record, and store information. Surgery is no exception to big data. New technologies have allowed researchers to collect intraoperative audiovisual recordings, capture instrument motion, or track surgeon's eye movements, that may prove to be useful in evaluating performance and ensuring patient safety.[98] In fact, more studies are demonstrating AI's aptitude in differentiating surgical maneuvers, detecting instruments, and predicting surgical skill from real surgical footage.[99-101]

Unfortunately, because most algorithms conducting intraoperative assessment utilize unsupervised deep learning methods, it is difficult to understand how the AI model is making its decision with our current technologies. This is known as the "Black Box" problem, and it is a major consideration in designing deep learning algorithms whose output decisions could have serious implications during high-stakes procedures like surgery.[102] Supervised deep learning is one way to make these algorithms more transparent. In this process a multilayered neural network is trained on data with a set of pre-defined features therefore providing stakeholders with a level of control on the criteria used for the algorithm's decision making.[98] One advantage of deep learning models is their ability to output real-time decisions using complex data.

To do so in the context of surgical performance assessment in simulation, our group recently developed the Intelligent Continuous Expertise Monitoring System (ICEMS).[8] This system uses a type of a recurrent neural network, known as a long short-term memory (LSTM) model, that evaluates timeseries input for ongoing assessment of expertise during procedures. The system makes an expertise prediction, that ranges from -1.00 (a "novice") to +1.00 (an "expert"), for every 0.2-seconds of the simulated procedure based on 16 performance features. These 16 metrics include the acceleration, velocity, force application, and the change in force application of each instrument, the instrument tip separation distance, the change in instrument tip separation distance, rates of brain and tumor resection, and four metrics associated with bleeding. The ICEMS has demonstrated ability to differentiate the performance of four groups (i.e., consultant surgeons and fellows, PGY4-6 senior residents, PGY1-3 junior residents, and medical students) based on their mean "Expertise Score"; the average of expertise predictions for the entire procedure.[8] This system is a granular, objective, and feasible tool to measure performance outcome in randomized trials that can potentially require the assessment of hundreds of procedures.

Intelligent Tutoring Systems

Intelligent tutoring systems are educational platforms that integrate assessment with personalized feedback.[13] In undergraduate medical education, there have been several intelligent tutoring systems developed to support learning clinical knowledge.[103] These systems have predominantly focused on the training of non-technical skills, such as medical students' diagnostic reasoning,[104] communication skills,[105] and clinical decision-making abilities.[106] The reason behind this is partly due to technological limitations in turning a surgical performance into data. Doing so requires generating enough data, from experts and novices on standardized procedures, so that AI algorithms can explore the composites of expert performance and elucidate clear learning objectives of the task.[103]

From an educational standpoint, the assessment rubric used by the algorithms need to contain explicit competency criteria that will ensure transparency in summative evaluation and can be used for formative feedback purposes.[23] Hence, supervised machine learning algorithms that use pre-defined features are also very useful. One problem with AI models, especially in supervised machine learning, is that they can be sensitive to the input data. In other words, too much variation in the surgical procedure, the tools involved, or the skills required may adversely influence their accuracy.[107] As a result, developing and testing new AI models holds promise in virtual reality simulation where the procedures are uniform, the data recordings are consistent, and the simulation environment is tightly controlled.[1] By collecting performance data on the subpial resection scenarios of the NeuroVR from expert consultants, residents, and medical students, our team identified measurable expert competence benchmarks.[7]

Having quantifiable benchmarks of expert performance, our team leveraged machine learning to design an intelligent tutoring system known as the Virtual Operative Assistant

(VOA).[16] This is an automated learning platform that uses raw data from the NeuroVR simulator to calculate learners' metric scores, classify their expertise (e.g., as either a novice or an expert), and determine whether their performance is within the expert benchmark for a given metric. If the learner is classified as a novice for any of the metrics, the system provides audiovisual feedback with specific actionable instructions to excel.[16]

VOA's predictive model is in the form of a linear support vector machine, that classifies learners' competence based on their metric scores. From roughly 6000 possible benchmark criteria, 270 relevant and teachable metrics were kept after expert consultation, and from these, four that most significantly determined an expert were selected by the support vector machine algorithm. These metrics were maximum bipolar force and mean bleeding rate (associated with safety), and average instrument separation distance and mean bipolar acceleration (associated with instrument movement).[16] Utilizing these metrics, VOA achieved an accuracy of 82% where all of the 28 experts (i.e., consultant surgeons, fellows, and senior residents in PGY4-6) were correctly classified as an "expert" but 4 of the 22 novices (i.e., junior residents in PGY1-3 and medical students) were incorrectly classified as "expert".[16]

VOA follows the competency-based model of the Royal College of Physicians and Surgeons of Canada and assesses individuals in two steps, first helping them reach competency in safety metrics (STEP-1) before allowing them to proceed to STEP-2 of the training for the evaluation of metrics associated with instrument movement.[16,108] This form of segmentation is thought to benefit junior learners and mitigate cognitive load.[54] By integrating personalized performance assessment and feedback, VOA is the first intelligent tutoring system for teaching psychomotor technical skills in surgical simulation. As such, we sought to test its effectiveness and compare it with the more conventional expert apprenticeship and instruction method.

Instructor Training

To best reproduce the apprenticeship learning experience for a controlled experiment during the COVID-19 pandemic posed some challenges, notably, that we needed to restrict contact and the number of individuals on site. Additionally, the presence of an expert instructor in the simulation room for one group, would introduce certain variables that may influence participants performance and emotions in a way that may not be possible to control for the other groups. Therefore, we decided to pursue remote instruction and use established assessment and debriefing methods in surgical instruction, namely the OSATS and the Promoting Excellence And Reflective Learning in Simulation (PEARLS) debriefing guide, respectively.[89,109] However, given the number of instruction hours required for a well-powered study, it was difficult to involve consultant neurosurgeons. Thus, we designed a three-week workshop and trained senior residents to take the role of an expert instructor. This training would also help to calibrate raters' criteria-based assessment, increase consistency, and ensure standardized feedback.[4,93]

Two senior neurosurgery residents (PGY5), with operative experience of the subpial procedure, received a structured training that prepared them for providing remote instruction. This three-week workshop followed principles of traditional expert apprenticeship supported by modules of independent deliberate practice guided by self-regulated learning. The resident's learning objectives were to 1) achieve expert competence in performing two subpial resection scenarios, 2) assess screen-recorded performance videos of students, and 3) provide debriefing and constructive feedback. To reduce bias in the instructors' assessment and instruction during the trial, they were blinded to the AI metrics during this training.

First, in two ninety-minute sessions of apprenticeship, residents observed and performed the practice and realistic simulated brain tumor removals, one scenario type per session, under the supervision of an expert. The goal here was to provide high-level coaching to the instructors in performing the scenarios and familiarizing them with standards of expertise used for assessment on the previously validated OSATS visual rating scale.[96] Then, each instructor had three one-hour sessions of deliberate practice with the explicit goal of improving performance to the expert level. Between simulations, instructors completed self-assessment of their screen-recorded performance to actively engage with their learning, make strategic choices that change their behaviour, and practice assessing videos using the OSATS scale. This fosters self-regulated learning strategies that enhances instructors' metacognitive processes and allows them to critically evaluate other student's performances.

Finally, in two one-hour sessions of "Transition to Instruction", instructors worked in pairs to provide peer assessment and feedback to each other, and rate pre-recorded on-screen performance of medical students separately for inter-rater reliability assessment. A feedback and debriefing script based on the PEARLS guide[109] was developed for instructors to follow during the training sessions with student participants. Feedback script is based on the OSATS assessment guide. It describes the lacking competency and the relevant OSATS category, and provides actionable instructions, given by senior consultants, on how to excel. Instructors had the freedom to rephrase statements or provide feedback outside the list, if necessary, in which case it was recorded and used future participants.

Assessment data was collected and analyzed for the OSATS scale's internal consistency and the evaluators' inter-rater reliability. Table 1 demonstrates that prior to study recruitment, the scale showed a good internal consistency (α=0.82, 95% CI 0.77-0.87) and instructors achieved good inter-rater reliability (ICC=0.84, 95% CI 0.79-0.88).

THE STUDY OBJECTIVES

Because the VOA is the first intelligent tutoring system in teaching surgery, its effectiveness compared with learning from a human expert is unknown. Therefore, this thesis aims to determine if the VOA is effective and compare it with remote expert instruction in 1) improving surgical performance, and 2) eliciting emotions and cognitive load. The primary research question was how do students learning with the VOA perform compared with those learning with a human mentor? The secondary research question was what are the affective and cognitive responses to these different modes of instruction?

THE STUDY HYPOTHESIS

Our hypothesis was that the VOA feedback would be non-inferior to remote expert instruction in performance outcomes, but it would lead to stronger negative emotions and higher cognitive load. Although studies comparing intelligent tutoring to expert instruction in learning technical skills are lacking, for conceptual knowledge, previous research comparing an intelligent algebra tutor to class instruction in high school students found no significant difference in the students' mathematics scores after one year.[110] Past findings have indicated that intelligent tutoring resulted in gradually stronger feelings of boredom, frustration, and confusion.[111,112] This, in addition to a higher number of interactive elements involved in virtual reality simulation as a result of VOA feedback, led us to hypothesize that learners may experience higher cognitive load.[55,113,114]

THE STUDY

Artificial Intelligence Tutoring Compared with Expert Instruction in Surgical Simulation Training. - A Randomized Controlled Trial

Ali M. Fazlollahi, BSc, Mohamad Bakhaidar, MD, Ahmad Alsayegh, MD, Recai Yilmaz, MD, Alexander Winkler-Schwartz, MD, Nykan Mirchi, MSc, Ian Langleben, Nicole Ledwos, MSc, Abdulrahman J. Sabbagh, MBChB, Khalid Bajunaid, MD, MSc, Jason M. Harley, PhD, Rolando F. Del Maestro, MD, PhD

Supplemental material is incorporated in the Appendix to give readers additional information on this work.

ABSTRACT

Importance: To better understand the emerging role of Artificial Intelligence (AI) in surgical training, efficacy of intelligent tutoring systems, such as the Virtual Operative Assistant (VOA), must be tested and compared to conventional educational approaches.

Objective: Determine how VOA compares with remote expert instruction in learners' performance, emotional, and cognitive outcomes during surgical simulation training.

Design: Instructor-blinded randomized controlled trial comparing VOA feedback with remote expert instruction and no-feedback controls. Cross-sectional data collected from January to April 2021.

Setting: McGill Neurosurgical Simulation and Artificial Intelligence Learning Centre, Montreal, Canada. Participants performed simulated surgical procedures with feedback delivered on site. Expert instructors observed participants' on-screen performance, provided live verbal feedback remotely.

Participants: Medical students from multiple institutions recruited through programs' social media and student networks.

Interventions: Five feedback sessions of 5 minutes each during a single 75-minute session, including 5 practice followed by 1 realistic virtual reality simulated brain tumor resections. Audiovisual metric-based feedback from VOA (VOA group) or verbal scripted debriefing and feedback from an instructor (instructor group) compared with control group receiving no performance feedback.

Main Outcome(s) and Measure(s): Primary 1) change in procedural performance, quantified as an "Expertise Score" by a validated deep learning assessment algorithm known as Intelligent Continuous Expertise Monitoring System (ICEMS) for each practice resection. 2) Learning and retention, measured as performance on realistic tumor resection by ICEMS and blinded Objective Structured Assessment of Technical Skills (OSATS). Secondary 1) strength of emotions before, during and after intervention, 2) cognitive load after intervention, measured by self-report questionnaires.

Results: Seventy medical students (41 female [%58.6]; mean [SD] age, 21.8 [2.3] years) from four institutions randomly assigned to VOA (n=23), instructor (n=24), and control (n=23) groups. All participants included in final intention-to-treat analysis. ICEMS assessed 350 practice resections, ICEMS and OSATS evaluated 70 realistic resections. During practice, VOA significantly improved Expertise Scores 0.66 (95% CI 0.55-0.77) and 0.65 (95% CI 0.54-0.77) points higher than instructor and control groups (p<.001). Realistic resections' Expertise Score was significantly higher for VOA group compared to instructor and control groups (mean difference 0.53; 0.49, respectively, p<.001). VOA and instructor group's realistic resection OSATS ratings were not significantly different. Compared to controls, VOA significantly enhanced overall score, expert instruction significantly improved instrument handling. No significant between-groups difference in cognitive load, positive activating, and negative emotions.

Conclusions and Relevance: VOA's quantitative benchmark feedback demonstrated superior performance outcome, improved skill transfer, non-inferior OSATS ratings, and equivalent cognitive and affective responses compared to remote expert instruction indicating advantages for its use in simulation training.

Trial Registration: ClinicalTrials.gov Identifier NCT04700384.

INTRODUCTION

Mastery of bimanual psychomotor skills is a defining goal of surgical education,[84,115] and wide variation in surgical skill among practitioners is associated with adverse intraoperative and postoperative patient outcomes.[4,5] Novel technologies, such as surgical simulators utilizing artificial intelligence (AI) assessment systems, are improving our understanding of the composites of surgical expertise and have the potential to reduce skill heterogeneity by complementing competency-based curriculum training.[1,68,116] Virtual reality simulation and machine learning algorithms can objectively quantify performance and improve precision and granularity of bimanual technical skills classification using expert benchmarks.[7,9,10] These systems may enhance surgical educators' ability to develop more quantitative formative and summative assessment tools to manage future challenging pedagogic requirements. The COVID-19 pandemic has significantly altered surgical trainees' ability to obtain intraoperative instruction necessary for skill acquisition,[117] and innovative solutions such as AI-powered tutoring systems may help in addressing such disruptions.[118]

An intelligent tutoring system refers to an educational platform driven by computer algorithms that integrate assessment with personalized feedback.[13] Our group has developed an intelligent tutoring system called the Virtual Operative Assistant (VOA) that utilizes a machine learning algorithm, support vector machine, to classify learner performance on validated benchmarks and provides goal-oriented, metric-based audiovisual feedback in virtual reality simulations.[16] Following the competency-based medical education model of the Royal College of Physicians and Surgeons of Canada,[86] and to mitigate extrinsic cognitive load through segmentation,[54] the system guides learners in two steps: First, helping trainees reach competency in safety metrics before evaluating metrics associated with instrument movement and efficiency.[16] The VOA intelligent tutoring system is designed for surgical simulation training, but its effectiveness compared to conventional surgical instruction is unknown.

Expert-led tele-mentoring and virtual clerkships use technologies such as augmented reality headsets and videotelephony softwares.[66,83] With the ongoing pandemic, these adaptations may provide alternatives to intraoperative surgical instruction.[119] For this study, we followed gold standards of assessment and debriefing in surgical education – Objective Structured Assessment of Technical Skills (OSATS)[89] and Promoting Excellence And Reflective Learning in Simulation (PEALRS) debriefing guide[109] – to design a standardized expert-led remote training as the traditional control.

We sought to investigate VOA's educational value by comparing it to remote expert instruction in enhancing technical performance and learning outcome of medical students during brain tumor resection simulations and eliciting emotional and cognitive responses that support learning. Our hypothesis was that VOA feedback would be non-inferior to remote expert instruction in performance outcomes but lead to stronger negative emotions and higher cognitive load.

METHODS

This multi-institutional instructor-blinded randomized controlled trial was approved by McGill University Health Centre Research Ethics Board, Neurosciences-Psychiatry, and registered on US National Library of Medicine (ClinicalTrials.gov, NCT04700384). This report follows Consolidated Standards of Reporting Trials involving AI[120] and Best Practices for Machine Learning to Assess Surgical Expertise.[12]

Participants

Medical students, with no surgical or virtual reality simulation experience were invited to voluntarily participate. Recruitment information was shared among student networks, social media, and interest groups. Selection was based on meeting inclusion criteria, enrollment in Medicine Preparatory, first- or second- year of a medical program in Canada, and not meeting the exclusion criteria, participation in surgical clerkship or previous experience with the NeuroVR (CAE Healthcare, Montreal, Canada) simulator. All participants signed an informed consent form prior to participation. Participant demographic information is outlined in Table 2.

Randomization

Students were stratified by gender and block randomized to three intervention arms using an internet-based computer-generated random sequence (random.org). Group allocation was concealed by study coordinator and instructors were notified of appointment times one day in advance for scheduling purposes. CONSORT flow diagram is outlined in Figure 4.

Study Procedure

After written consent, a background information questionnaire was administered that recorded baseline emotions, experiences that may influence bimanual dexterity (video games,[121] musical instruments[122]), deliberate practice (competitive sports[123]), or prior virtual reality navigation. Students were not informed of the trial purpose or assessment metrics. Participants performed 5 practice simulated tumor resections (Figure 2),[73] followed by feedback or no-feedback (control), then 1 realistic tumor resection simulation (Figure 3)[72] to evaluate learning and skills transfer. Two self-report questionnaires administered upon completion of the fifth and sixth resections assessed participants' emotions during and after the learning session, respectively, and measured cognitive load after training.

Simulator

The NeuroVR simulates neurosurgical procedures on a high-fidelity platform that recreates the visual, auditory, and haptic experience of resecting human brain tumors (Figure 1).[6] Because this simulator records timeseries data of users' interaction in the virtual space,[71] machine learning algorithms have been demonstrated to successfully differentiate surgical expertise based on validated performance metrics.[7,8,10]

Virtual Reality Tumor Resection Procedures

Subpial resection is a neurosurgical procedure in oncologic and epilepsy surgery requiring coordinated bimanual psychomotor ability to resect pathologic tissue with preservation of surrounding brain and vessels.[3] Students' objective was to remove a simulated cortical tumor with minimal bleeding and damage to surrounding tissues using a simulated aspirator in the dominant hand and a simulated bipolar forceps in the non-dominant hand, for manipulating tissues and cauterizing bleeding.[72,73] Participants received standardized verbal and written instructions on instrument use and performed orientation modules to understand each instrument's functions. Individuals had 5 minutes to complete each practice resection and 13 minutes for the realistic resection. The first practice subpial resection was considered baseline performance.

Interventions

Five minutes were allowed for intervention between practice resections. Both experimental arms follow principles of deliberate practice guided by self-regulated learning (SRL),[32,124] where formative assessment enables finding areas of growth, setting goals, and adopting strategies that enhance competence.[31] Feedback students received and progress towards learning objectives was monitored by either the VOA or an instructor.

Virtual Operative Assistant Intelligent Tutoring

VOA predicts a competence percentage score based on four metrics: assessment criteria selected through expert consultation, statistical, forward, and backward support vector machine feature selection.[16] Competence is evaluated in two steps, safety (STEP-1) and instrument movement (STEP-2), each associated with two metrics: mean bleeding rate and maximum bipolar force application for STEP-1, average instrument tip separation distance and mean bipolar acceleration for STEP-2. Learners must achieve "expert" classification for safety metrics in STEP-1 before moving to STEP-2 to learn instrument movement metrics and achieve competency. Individuals classified as "novice" in any metric, receive automated audiovisual feedback (Figure 5).[16]

Remote Expert Instruction

Delivering traditional apprenticeship learning during the COVID-19 pandemic for a controlled experiment requires steps that minimize contact and ensure consistency. Two senior neurosurgery residents (M.B., A.A., post-graduate year 5) who had experience performing human subpial resection procedures, completed standardized training (Appendix 1) to perform simulations within consultants' benchmarks, reliably rate on-screen performances using a modified OSATS visual rating scale,[96] and provide feedback from a modified PEARLS debriefing script.[109] Instructors were blinded to AI assessment metrics. Prior to recruitment, OSATS scale demonstrated good internal consistency (α=0.82, 95% CI 0.77-0.87) and instructors achieved good inter-rater reliability (ICC=0.84, 95% CI 0.79-0.88).

Each participant's live on-screen practice performance was assessed remotely by one randomly selected instructor who completed an Assessment Sheet (Figure 6, Appendix 2). During debriefing, instructors followed a modified PEARLS script and provided feedback from a list of instructions, suggested by consultants, depending on students' competency. Table 3 contains details on feedback interventions.

Control Group

Control participants received no performance assessment or feedback and were instructed to use the time between simulations to reflect and set goals for the following trial. This follows principles of experiential learning through active experimentation and reflective observation,[125] establishing a baseline for performance improvement and learning with no feedback.

Outcome Measures

Primary outcome was interaction effect of feedback on surgical performance improvement over time during 5 practice resections, measured by the Intelligent Continuous Expertise Monitoring System's (ICEMS) Expertise Score: average of expertise predictions (range, -1.00 to +1.00, reflecting "novice" and "expert", respectively) computed for every 0.2-second of the procedure, by a deep learning algorithm utilizing a long short-term memory

(LSTM) network, using 16 metrics and simulator's raw data as input.[8] Learning and skill retention was evaluated based on realistic tumor resection performance by both the ICEMS and blinded OSATS assessment. Secondary outcomes were differences in strength of emotions before, during, and after training, and cognitive demands required by each intervention, measured by Duffy's Medical Emotions Scale,[44] and Leppink's Cognitive Load Index,[53] respectively, using self-report questionnaires.

Statistics

Ad-hoc analysis to achieve 80% statistical power (β=0.20), estimating moderate primary outcome effect of 35%, with two-sided test at α=0.05, revealed minimum of 23 participants required for each intervention arm. Collected data was examined for outliers and normality. Levene's test for equality of variance and Mauchly's test of sphericity met assumptions of analysis of variance. Two-way mixed ANOVA investigated interaction of group assignment (between-subjects) and time (within-subjects) on learning curves and emotion self-reports. One-way ANOVA tested between-group differences in learning, cognitive load, and OSATS scores. Baseline performance was assigned as covariate in the mixed model. Repeated measures ANOVA examined within-subjects changes of performance in each group. Significance was set at p<.05. All statistical analyses were performed on SPSS Version 27 (IBM Corporation, Version 27). Expertise Score predictions were conducted in MATLAB (MathWorks Inc. 2020a release).

RESULTS

Seventy medical students (41 female [58.6%]; mean [SD] age, 21.8 [2.3] years) from four institutions (McGill University, 32 [45.7%]; Laval University, 19 [27.1%], University of

Montreal, 17 [24.3%], University of Sherbrooke, 2 [2.9%]) were randomly assigned to either VOA (n=23), instructor (n=24) or control (n=23) groups. Between-group distribution of baseline characteristics was balanced. Three-hundred and fifty practice and 70 realistic resections were scored by the ICEMS. Blinded experts evaluated 70 video recordings of realistic performances using the OSATS scale. There was no between-groups baseline performance difference. All VOA group participants passed the safety module (STEP-1) and 14 (61%) completed instrument movements competency (STEP-2) by the end of training (Figure 7).

Performance During Practice Tumor Subpial Resection

Mixed ANOVA demonstrated that within-subjects performance changes depended on the type of feedback. VOA's feedback group achieved 0.66 (95% CI 0.55-0.77, p<.001) and 0.65 (95% CI 0.54-0.77, p<.001) points higher Expertise Scores than instructor and control groups, respectively (Figure 8A). Mean Expertise Scores in instructor and control groups were not significantly different.

VOA group demonstrated Expertise Scores improvements between trials (Figure 8A). Pairwise comparisons demonstrated that learners performed significantly better than baseline after intelligent tutoring feedback (mean difference 0.37, 95% CI 0.18-0.56, p<.001, with Trial-1; 0.51, 95% CI 0.29-0.74, p<.001, with Trial-2; 0.65, 95% CI 0.41-0.89, p<.001 with Trial-3; 0.61, 95% CI 0.36-0.86, p<.001 with Trial-4). There was significant improvement from Trial-1 to Trial-3 (0.28, 95% CI 0.55-0.02, p=.02) and Trial-4 (0.24, 95% CI 0.00-0.49, p=.04). Learning curves demonstrate steady improvement from baseline to Trial-3 that plateaus at Trial-3 and 4. Three VOA feedback instances resulted in average group performance above 0.00: ICEMS's "novice" threshold (Figure 8A).

Three of 4 VOA metrics used for competency training, maximum bipolar force application, average instrument tip separation distance and mean bipolar acceleration, demonstrated improvement in VOA group and significant differences to instructor and control groups (Figure 8B-D). There was no significant between-groups difference in bleeding rate due to wide participant variability in this metric. Although VOA feedback was more effective in enhancing metric scores compared to expert instruction, compared to control, remote expert feedback significantly reduced average instrument tip separation distance (mean difference -3.28, 95% CI -6.36 to -0.21, p=.034) (Figure 8C). Eight of the 16 ICEMS metrics not trained by the VOA had significantly improved in the VOA group compared to instructor and control conditions (results not shown) suggesting that feedback on 4 AI-selected safety and instrument movement metrics resulted in improved bimanual psychomotor performance in other benchmark metrics.

Realistic Tumor Resection Performance

VOA group achieved significantly higher Expertise Scores in the realistic subpial resection than instructor and control groups (mean difference 0.49, 95% CI 0.34-0.61, p<.001; 0.53, 95% CI 0.40-0.67, p<.001, respectively) (Figure 9A). Realistic subpial resection's global OSATS ratings showed no significant difference between VOA and instructor groups, consistent with an equivalent qualitative performance outcome. Compared to control group, feedback significantly improved participants' respect for tissue (mean difference 1.17, 95% CI 0.40-1.95, p=.002, VOA; 0.85, 95% CI 0.08-1.62, p=.027, instructor) and economy of movement (mean difference 1.35, 95% CI 0.39-2.31, p=.004, VOA; 1.07, 95% CI 0.12-2.02, p=.024, instructor), while expert instruction significantly enhanced instrument handling (mean difference 1.18, 95%

CI 0.22-2.14, p=.012), VOA resulted in significantly higher overall scores (mean difference 1.04, 95% CI 0.13-1.96, p=.021) (Figure 9C). Completing VOA's instrument movement competency correlated significantly with higher economy of movement (Pearson coefficient 0.25, p=.03) suggesting successful acquisition of the relevant competency.

Emotions and Cognitive Load

Within-subjects, there was a significant increase in positive activating emotions (mean difference after-before 0.36, 95% CI 0.16-0.55, p<.001) and a significant decline in negative activating emotions (mean difference after-before -0.59, 95% CI -0.85 to -0.34, p<.001) throughout the simulation training (Figure 10A-C). Significant interaction effect in positive deactivating emotions demonstrated that instructor group participants felt more relieved and relaxed during training compared to learners in VOA and control groups (mean difference 0.75, 95% CI 0.19-1.31, p=.01; 0.71, 95% CI 0.14-1.27, p=.01, respectively). No between-subjects difference in intrinsic, extrinsic, and germane cognitive load were found (Figure 10D).

DISCUSSION

Surgical performance is an independent predictor of postoperative patient outcomes[91] and technical skills acquired in simulation training improve operating room performance.[126-128] Repetitive practice in a controlled environment and educational feedback are key features of simulation-based surgical education,[129] however, use of autonomous pedagogical tools in simulation training is limited. To our knowledge, this is the first study that compares effectiveness of an AI-powered intelligent tutoring system with expert instruction in surgical simulation while assessing affective and cognitive response to such instruction.

In this randomized controlled trial, our findings demonstrated effective use of intelligent tutoring in surgical simulation training. VOA feedback improved performance during practice and realistic simulation scenarios, measured quantitatively by Expertise Scores, and enhanced operative quality and students' skill transfer, observed by OSATS during the realistic tumor resection. Objective metric-based formative feedback through intelligent tutoring demonstrated advantages compared to remote expert instruction. It helped students achieve higher expertise by bringing awareness to their metric goals during resections and setting measurable performance objectives, two effective strategies of learning theory.[35] Feedback on AI-selected metrics had an extended effect on supplementary performance criteria used in both OSATS and ICEMS. VOA's learning platform is flexible and allows learners with different levels of expertise to practice and receive personalized formative feedback based on interest and time availability. VOA did not bring participants' Expertise Scores to the level of senior experts (ICEMS > 0.33)[8] suggesting areas for future research and improvement.

In contrast to our hypothesis and previous reports, where learning with an intelligent tutor elicited negative emotions, impairing students' use of SRL strategies,[111,112] learning bimanual tumor resection skills with VOA demonstrates a gradual decline in negative activating emotions with an overall increase in positive emotions, comparable to human instruction. Encouragingly, VOA participants did not report this learning experience required significantly higher cognitive demands compared to the other interventions, demonstrating clear and comprehensible intelligent tutoring feedback that required minimal extraneous load.

Although the full-impact of COVID-19 on surgical education remains unclear,[130] it is important to prepare for future challenges through focused research and further development of effective remote learning platforms.[76] We report two potential methods to address remote learning with the goal of enhancing task performance, from which, intelligent tutoring is more efficient and effective. Similar to previous studies,[131,132] our findings suggest that scripted feedback by instructors established a supportive learning environment where participants felt stronger positive deactivating emotions during practice, however, this did not result in greater performance. Studies suggest that there is no statistically significant difference in complication rates, operative time, and surgical outcomes between tele-mentoring and in-person instruction,[133,134] but there is limited evidence comparing their educational effectiveness on technical performance. In this study, remote instruction was inferior to intelligent tutoring, based on quantifiable metrics but further research is necessary to determine if that remains the case with in-person coaching. Our remote-based method was considered feasible by instructors because they could easily join to provide virtual debriefing and technical instruction.

The AI algorithm utilized in this study failed to detect instructor group's performance improvements according to OSATS ratings for practice and realistic scenarios (Figure 11). OSATS categories like instrument handling, describe a subjective qualitative composite of actions that AI systems have difficulty measuring from raw data. ICEMS functions at a deeper level by analyzing the interaction of several underlying metrics that contribute to expertise. These systems are less able to assess operative strategies, such as a systematically organized tumor resection plan, that students may acquire more readily from expert instruction. These types of procedural instruction may take more educator time to become apparent as changes in learners' metrics scores. Our findings suggest that monitoring specific AI-derived expert performance metrics, such as bipolar instrument's acceleration and providing personalized quantitative learner feedback on these metrics, is an efficient methodology to guide behavioral changes towards a higher operative quality. However, integrating metric objectives with the task goals may be challenging and require expert input. Most participants (93%) reported that they would prefer learning with feedback from both expert instruction and intelligent tutoring, suggesting complementary features from both methods could enhance the learning experience. With increasing efforts to capture live operative data,[98] combining intraoperative use of intelligent tutoring and expert surgical instruction may accelerate the path to mastery.

Limitations

Although this AI-powered virtual reality simulation platform allows detailed quantitative assessment of bimanual technical skills, it fails to capture the full spectrum of competencies required in surgery. Other limitations include the sample cohort, use of volunteers, instructor experience level, and the remote instruction context that limited in-person expert feedback delivery due to the pandemic. Future studies should focus on evaluating the effectiveness of intelligent tutoring systems compared with both remote and in-person expert instruction on learner simulated surgical performance. Investigations combining intelligent tutoring with personalized expert feedback and debriefing may also increase our understanding of the value of these technologies.

Conclusion

Performing simulated brain tumor resections was more effective with feedback from an intelligent tutor compared with learning from remote expert instruction. VOA significantly improved Expertise and OSATS scores in a realistic procedure while fostering an equivalent affective and cognitive learning environment.

THESIS SUMMARY

Discussion

To test the effectiveness of an intelligent tutoring system in neurosurgical simulation training, a parallel design instructor blinded randomized controlled trial was conducted with 70 medical student participants from four Canadian institutions. Participants' technical performance improvement, skill acquisition and transfer, emotions, and cognitive load were measured in three intervention arms.

Foremost, VOA's feedback was successful in directing behavioural actions that changed participants' scores on VOA's intrinsic metrics. Three of the four VOA metrics, maximum bipolar force application, average instrument tip separation distance, and mean bipolar acceleration, significantly decreased from baseline after VOA feedback (Figure 8C-D). Furthermore, all individuals receiving the VOA training passed the safety competency and a majority (61%) completed the instrument movement competency as well (Figure 7). The competencies acquired during the five practice resections were transferred to the realistic tumor resection scenario and observed by blinded instructors on the OSATS scale (Figure 9C). A significant positive correlation between completion of VOA's instrument movement competency and the OSTAT's economy of movement scores further demonstrated effective learning and skill transfer from metric-based assessment and feedback.

An important question raised by scholars is if intelligent tutoring feedback results in improved performance only in the metrics measured by the system, or whether it yields to performance changes in other areas.[135] In this study, VOA improved performance of medical students in three OSATS categories of respect for tissue, economy of movement, and the overall quality (Figure 9C). By examining all 16 metrics of the ICEMS, we also found improved performance after VOA feedback in eight other metrics not included in the VOA. From these metrics, four were directly associated with the VOA feedback. These were average bipolar force application, average change of bipolar force application, the velocity of the bipolar, and the change in instrument tip separation distance. VOA's feedback could have a conceivable direct effect in these metrics. For example, reducing the maximum force applied with the bipolar will directly affect the average bipolar force application. However, we identified four extrinsic metrics that demonstrated performance improvement in the VOA group despite having no direct or conceivable immediate relation with the VOA feedback. These included the rates of tumor and healthy tissue removal, and the acceleration and velocity of the aspirator.

We observed changes in the movement of the dominant hand (acceleration and velocity of the aspirator) due to feedback on the metrics associated with the non-dominant hand which is a topic of investigation by our team. This inter-manual transfer of skills was also observed in a different randomized controlled trial involving non-dominant hand training in laparoscopic surgery.[136] Functional neuroimaging may provide some answers to the unknown mechanism behind these observations. A study investigating brain activation using functional near-infrared spectroscopy (fNIRS) demonstrated that the use of chopsticks in the left hand of right-handed individuals led to bilateral activation of the motor cortex and premotor area, whereas, its use in the dominant hand only activated the contralateral primary motor area (M1).[137] This suggests that perhaps the awareness required to control non-dominant hand movements, results in a change of movements of the dominant hand due to functional dominance and network asymmetry between the two hemispheres.[137,138] This is interesting considering that three of VOA's four AI-selected metrics assess learners' use of their non-dominant hand in this bimanually complex surgical task, and our results show that practicing on these metrics is an efficient method of training.

Student's learning through remote instruction did not improve as well as the VOA group in the same metrics. But they improved significantly better than the control group in instrument tip separation distance (Figure 8C) and the rate of healthy tissue removal, two of the 16 ICEMS performance metrics. However, because they were instructed to use the bipolar more frequently, their force metrics of the non-dominant hand were significantly higher, thus more novice, than the control participants (not shown). This is perhaps another reason why there was no statistically significant difference in the Expertise Score between these groups. Instructor group participants achieved a higher OSATS score in respect for tissue, economy of movement, and instrument handling in the realistic tumor resection compared to the no-feedback control (Figure 9C). This enhanced quality was not detected by the ICEMS, but it demonstrated that remote-instruction and scripted feedback were effective in enhancing aspects of the operative quality. This is still a significant finding because instructors in this study were limited in formal teaching experience and instructional freedom such as the use of visual demonstrations. Randomized trials comparing VOA training with in-person coaching are necessary to perform and they will be considered when it is safe and feasible to do so.

The remote-instruction sessions were reported to be feasible by the instructors as they could join the sessions virtually from anywhere with an internet connection and a computer device. Our methods outline a systematic approach to train post-graduate medical trainees to reliably use educationally relevant tools, such as criteria-based assessment rubrics and structured debriefing guides, to effectively train undergraduate medical students. This may prove to be a useful tool in remote and under-resourced areas where the proportion of available expert instructors may be fewer than 1 neurosurgeon per several million people.[139] Previous studies utilizing tele-mentoring in neurosurgery have also shown to be feasible in enhancing anatomical understanding of medical students or guiding residents in exposing the anterior circulation of a cadaver.[140] However, because access to expensive virtual reality simulators, neurosurgical equipment, and human cadaver facilities is likely to be limited in remote areas for training, future studies are required to assess whether similar technical performance outcomes will be observed in more accessible simulation models.

The cognitive load findings of this study demonstrated that the type of feedback did not affect learners' perception of the task complexity (i.e., intrinsic load), or their cognitive processing devoted to learning the task (i.e., germane load) (Figure 10D). Our sample was a motivated cohort of medical student volunteers who had expressed a high interest to pursue a surgical specialty training in the future (mean interest [SD], 3.80 [1.10] out of 5), Table 2. Participants attending this training were eager to learn about neurosurgery and gain technical experience, but they were relatively novice and had no surgical experience which may explain the high germane load reported by all three groups. A more experienced cohort with well-established mental models of surgery and prior neurosurgical simulation exposure may report that learning this procedure requires less internal cognitive resources.

Intrinsic load findings show that this surgical task had a moderate difficulty for medical students and there was a significant positive correlation (Pearson coefficient 0.39, p=.001) between strength of negative activating emotions during training and intrinsic load. Cognitive load imposed for comprehending feedback by both the VOA and instructors was surprisingly low. As described previously, highly immersive simulation training for novice learners is generally thought to increase cognitive load for novice learners.[56] We had hypothesized that adding intelligent tutoring to the simulation training would further increase the interactive elements involved and thus lead to higher extrinsic load.[113] But VOA's feedback component did not significantly increase the extrinsic cognitive load compared to no-feedback in the control group and remote-instruction by experts. This also suggests that medical students required a similar level of cognitive processing to comprehend instructions from the intelligent tutor as they did for understanding the instructions from experts. This is also promising because learners could effectively use the system independently for deliberate practice of basic skills.

Emotions elicited by the VOA platform were generally positive and equivalent to the control and instructor groups. All groups reported a significant increase in positive activating emotions (i.e., hope, gratitude, and happiness) and a significant decline in negative activating emotions (i.e., anxiety and confusion) with no significant between-subjects difference. The similarity in individuals' emotional responses to positive and negative activating emotions indicates that any performance difference observed between groups was likely the result of the instructional component of feedback. It is important to highlight that the VOA does not register learners' emotional cues and is not designed to provide psychological support. Embedding emotional state reasoning in intelligent tutoring systems has been attempted in the past and they could be considered for VOA's future improvements.[141] To do so, VOA's feedback model will need to incorporate two supporting algorithms; a classifier that can accurately predict learners' affective state and a decision-tree to personalize the metric-specific feedback with effective emotion regulating strategies. This can be attempted using Emotion Regulation in Achievement Situations (ERAS) model as the theoretical framework for the decision-tree and learners' skin conductance rate as input for the emotions classifier.[43,46] However, how much benefit this would add to the VOA remains to be investigated.

A previous study comparing Affective AutoTutor, an emotion-sensitive intelligent tutor for teaching introductory physics concepts, with regular AutoTutor found no significant learning gains between the two systems, but it also demonstrated that the animated pedagogical agent was more effective only for students with low-domain knowledge.[142] Similar results were found in a separate study that showed despite the affect-aware intelligent tutor's ability to reduce boredom and off-task behaviour, it did not result in a significant post-test performance difference compared to control.[143] Enhancing engagement is important for learning conceptual knowledge, but for hands-on technical activities, such as this surgical simulation training, affect-aware intelligent tutors are unlikely to further enhance engagement. Post-graduate learners who would use the VOA will have a higher domain knowledge than medical students, therefore the effect of emotion-sensitive feedback for them is likely to be small.

Debriefing by an expert instructor is likely to be a more effective approach to benefit the VOA platform. As demonstrated by our results, remote instruction resulted in learners feeling significantly more relaxed and relieved during the simulation training compared to the VOA condition and it was better than the VOA in significantly improving instrument handling performance compared to control. Research on structured debriefing in medical education has demonstrated many benefits for mastery learning.[131] This type of learner-centered coaching helps students become more reflective on their performance and fosters a psychologically safe learning environment that helps learners accept challenges.[144] As indicated above, instructors may provide helpful strategies to align the objectives of the task with VOA's metric goals to ensure learners develop the appropriate techniques. Future trials that combine complementary features from both methods are warranted. A blended approach that integrates VOA's metric-based assessment and feedback with structured debriefing and instruction from experts is likely to further enhance the performance and Expertise Score attainable by medical students.

Limitations and Future Directions

It must be noted that this virtual reality simulation training focused on only one element of surgical competence, which is the technical skills required to perform a sensitive step of a procedure involving multiple stages. As such, learners' expertise was measured solely based on their psychomotor ability to perform this task within experts' benchmarks. True surgical expertise is multifaceted and involves several other competencies, including team communication, leadership, and clinical reasoning. Therefore, this report's findings regarding the effectiveness of the VOA are limited to technical skill acquisition in virtual reality simulation.

The sample cohort in this study involved motivated medical student volunteers which increases the chance of bias, especially in self-report responses. Because this mode of simulation-based learning is most likely to be used for surgical training of post-graduate trainees this study's cohort limits the generalizability of our findings. Future trials must involve surgical residents, but the main limitation with that would be recruiting a sufficient sample size. With only a few neurosurgical residents per institution, a collective effort is required by the training programs to accrue enough individual for a well-powered study. Even then, this sample may be too small to yield sufficient power if it were to be randomized to more than two groups. Therefore, studies involving medical students can continue to inform us on effective instructional elements of intelligent tutoring before they could be incorporated as an appropriate intervention for residents.

Remote instruction by trained residents is limited in both the formal teaching experience of the instructors and their ability to demonstrate technical information. Our results suggest that for a medical student sample, trained residents can provide effective instruction. However, expert neurosurgeon consultants with extensive educational experience would be more appropriate for a resident population. It is important for future studies to compare the effect of in-person coaching alone with intelligent tutoring supplemented by either in-person or remote-instruction. Because instructors in this study were blinded to the assessment metrics used by the VOA, investigating the combined effect of in-person instruction with VOA feedback in this study was not feasible.

While other simulation-based training studies have been able to demonstrate the transfer of skills from virtual reality to the operating room or to improve procedural performance on cadavers,[126,128] doing so remains a challenge in simulations involving brain tumor surgery. The human brain is a soft tissue that significantly loses its natural consistency following preservation. Ex-vivo animal brains with gel-based tumors offer a suitable alternative as they can be acquired fresh and prepared for training.[59] This would help in testing the transferability of skills learned from the VOA to a controlled laboratory environment. To obtain information on the effect of AI feedback on real operative performance, a training program needs to be designed where some learners undergo virtual reality training with AI tutoring and in-person instruction while others receive virtual reality training with only in-person instruction. This longitudinal approach would also enable researchers to track the complication rates and postoperative outcomes of patients to see if this feedback can lead to a meaningful effect in patients' quality of life.

REFERENCES

1. Rogers MP, DeSantis AJ, Janjua H, Barry TM, Kuo PC. The future surgical training paradigm: Virtual reality and machine learning in surgical education. *Surgery.* 2021;169(5):1250-1252.
2. Dewan MC, Rattani A, Fieggen G, et al. Global neurosurgery: the current capacity and deficit in the provision of essential neurosurgical care. Executive Summary of the Global Neurosurgery Initiative at the Program in Global Surgery and Social Change. *Journal of Neurosurgery JNS.* 2019;130(4):1055-1064.
3. Hebb AO, Yang T, Silbergeld DL. The sub-pial resection technique for intrinsic tumor surgery. *Surgical neurology international.* 2011;2.
4. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical Skill and Complication Rates after Bariatric Surgery. *New England Journal of Medicine.* 2013;369(15):1434-1442.
5. Stulberg JJ, Huang R, Kreutzer L, et al. Association Between Surgeon Technical Skills and Patient Outcomes. *JAMA Surgery.* 2020;155(10):960-968.
6. Delorme S, Laroche D, DiRaddo R, Del Maestro RF. NeuroTouch: a physics-based virtual simulator for cranial microneurosurgery training. *Operative Neurosurgery.* 2012;71(suppl_1):ons32-ons42.
7. Winkler-Schwartz A, Yilmaz R, Mirchi N, et al. Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Network Open.* 2019;2(8):e198363-e198363.
8. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Del Maestro R. O51: ARTIFICIAL INTELLIGENCE UTILIZING RECURRENT NEURAL NETWORKS TO CONTINUOUSLY MONITOR COMPOSITES OF SURGICAL EXPERTISE. *British Journal of Surgery.* 2021;108(Supplement_1):znab117. 051.
9. Siyar S, Azarnoush H, Rashidi S, et al. Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Medical & Biological Engineering & Computing.* 2020;58(6):1357-1367.
10. Bissonnette V, Mirchi N, Ledwos N, et al. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *JBJS.* 2019;101(23).
11. Mirchi N, Bissonnette V, Ledwos N, et al. Artificial Neural Networks to Assess Virtual Reality Anterior Cervical Discectomy Performance. *Operative Neurosurgery.* 2020;19(1):65-75.
12. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial Intelligence in Medical Education: Best Practices Using Machine Learning to Assess Surgical Expertise in Virtual Reality Simulation. *Journal of Surgical Education.* 2019.
13. Ma W, Adesope OO, Nesbit JC, Liu Q. Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology.* 2014;106(4):901-918.
14. Penalosa E, Castaneda S. Meta-Tutor: an online environment for knowledge construction and self-regulated learning in clinical psychology teaching. *International Journal of Continuing Engineering Education and Life Long Learning.* 2008;18(3):283-297.
15. D'Mello S, Lehman B, Graesser A. A Motivationally Supportive Affect-Sensitive AutoTutor. 2011.
16. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLOS ONE.* 2020;15(2):e0229596.

17. RiskAnalytica. *The Case for Investing in Patient Safety in Canada.* August 2017.

18. Iyengar JV, Hargett GH, Hargett JD. Military development and applications of simulation systems. *Journal of International Information Management.* 1999;8(2):5.

19. Lee AT. *Flight simulation: virtual environments in aviation.* Routledge; 2017.

20. Hill RR, Tolk A. A History of Military Computer Simulation. In: Tolk A, Fowler J, Shao G, Yücesan E, eds. *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences.* Cham: Springer International Publishing; 2017:277-299.

21. Pedowitz RA, Marsh LJ. Motor Skills Training in Orthopaedic Surgery: A Paradigm Shift Toward a Simulation-based Educational Curriculum. *JAAOS - Journal of the American Academy of Orthopaedic Surgeons.* 2012;20(7).

22. Aggarwal R. Simulation in Surgical Education. In: Nestel D, Dalrymple K, Paige JT, Aggarwal R, eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Singapore: Springer Singapore; 2019:269-278.

23. Szasz P, Grantcharov TP. The Role of Assessment in Surgical Education. In: Nestel D, Dalrymple K, Paige JT, Aggarwal R, eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Singapore: Springer Singapore; 2019:221-228.

24. Abbas JR, Kenth JJ, Bruce IA. The role of virtual reality in the changing landscape of surgical training. *The Journal of Laryngology & Otology.* 2020;134(10):863-866.

25. Duvivier RJ, van Dalen J, Muijtjens AM, Moulaert VRMP, van der Vleuten CPM, Scherpbier AJJA. The role of deliberate practice in the acquisition of clinical skills. *BMC Med Educ.* 2011;11(1):101.

26. Ericsson KA, Krampe RT, Tesch-Römer C. The role of deliberate practice in the acquisition of expert performance. *Psychological review.* 1993;100(3):363.

27. Moonesinghe SR, Lowery J, Shahi N, Millen A, Beard JD. Impact of reduction in working hours for doctors in training on postgraduate medical education and patients' outcomes: systematic review. *Bmj.* 2011;342:d1580.

28. Gelfand DV, Podnos YD, Carmichael JC, Saltzman DJ, Wilson SE, Williams RA. Effect of the 80-Hour Workweek on Resident Burnout. *Archives of Surgery.* 2004;139(9):933-940.

29. Jarman BT, Miller MR, Brown RS, et al. The 80-hour work week: will we have less-experienced graduating surgeons? *Curr Surg.* 2004;61(6):612-615.

30. Dehmer GJ, Holper EM. Does practice make perfect? In: American College of Cardiology Foundation Washington DC; 2017.

31. Ericsson KA, Hoffman RR, Kozbelt A, Williams AM. *The Cambridge handbook of expertise and expert performance.* Cambridge University Press; 2018.

32. Zimmerman BJ. Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal.* 2008;45(1):166-183.

33. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Prentice-Hall; 1986.

34. Brydges R, Manzone J, Shanks D, et al. Self-regulated learning in simulation-based training: a systematic review and meta-analysis. *Medical Education.* 2015;49(4):368-378.

35. Kaufman DM, Mann KV. Teaching and Learning in Medical Education: How Theory can Inform Practice. In: *Understanding Medical Education.*2010:16-36.

36. Cruess RL, Cruess SR, Steinert Y. Medicine as a Community of Practice: Implications for Medical Education. *Academic Medicine.* 2018;93(2).
37. Rosen MA, Hunt EA, Pronovost PJ, Federowicz MA, Weaver SJ. In situ simulation in continuing education for the health care professions: A systematic review. *Journal of Continuing Education in the Health Professions.* 2012;32(4):243-254.
38. Pekrun R. The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators. *Applied Psychology.* 1992;41(4):359-376.
39. Brasil GDC, Lima LTB, Cunha EC, Cruz F, Ribeiro LM. Stress level experienced by participants in realistic simulation: a systematic review. *Rev Bras Enferm.* 2021;74(4):e20201151.
40. Pekrun R, Lichtenfeld S, Marsh HW, Murayama K, Goetz T. Achievement Emotions and Academic Performance: Longitudinal Models of Reciprocal Effects. *Child Dev.* 2017;88(5):1653-1670.
41. Bajunaid K, Mullah MAS, Winkler-Schwartz A, et al. Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. *Journal of Neurosurgery JNS.* 2017;126(1):71-80.
42. Pekrun R, Perry RP. Control-value theory of achievement emotions. In: *International handbook of emotions in education.* New York, NY, US: Routledge/Taylor & Francis Group; 2014:120-141.
43. Harley JM, Jarrell A, Lajoie SP. Emotion regulation tendencies, achievement emotions, and physiological arousal in a medical diagnostic reasoning simulation. *Instructional Science.* 2019;47(2):151-180.
44. Duffy MC, Lajoie SP, Pekrun R, Lachapelle K. Emotions in medical education: Examining the validity of the Medical Emotion Scale (MES) across authentic medical learning environments. *Learning and Instruction.* 2020;70:101150.
45. Goetz T, Hall NC. Emotion and achievement in the classroom. In: *International guide to student achievement.* New York, NY, US: Routledge/Taylor & Francis Group; 2013:192-195.
46. Harley JM, Pekrun R, Taxer JL, Gross JJ. Emotion Regulation in Achievement Situations: An Integrated Model. *Educational Psychologist.* 2019;54(2):106-126.
47. Schunk DH. *Learning theories an educational perspective sixth edition.* Pearson; 2012.
48. Paivio A. *Mental representations: A dual coding approach.* Oxford University Press; 1990.
49. Baddeley A. *Working memory, thought, and action.* Vol 45: OuP Oxford; 2007.
50. Sweller J. Cognitive load during problem solving: Effects on learning. *Cognitive science.* 1988;12(2):257-285.
51. van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. *Med Educ.* 2010;44(1):85-93.
52. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* 2014;36(5):371-384.
53. Leppink J, Paas F, Van der Vleuten CPM, Van Gog T, Van Merriënboer JJG. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods.* 2013;45(4):1058-1072.
54. van Merriënboer JJG, Kester L. The Four-Component Instructional Design Model: Multimedia Principles in Environments for Complex Learning. In: Mayer RE, ed. *The*

Cambridge Handbook of Multimedia Learning. 2 ed. Cambridge: Cambridge University Press; 2014:104-148.

55. Haji FA, Cheung JJH, Woods N, Regehr G, de Ribaupierre S, Dubrowski A. Thrive or overload? The effect of task complexity on novices' simulation-based learning. *Medical Education.* 2016;50(9):955-968.
56. LaRochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten C, Schuwirth L. Impact of increased authenticity in instructional format on preclerkship students' performance: a two-year, prospective, randomized study. *Acad Med.* 2012;87(10):1341-1347.
57. Calio BP, Kepler CK, Koerner JD, Rihn JA, Millhouse P, Radcliff KE. Outcome of a resident spine surgical skills training program. *Clinical spine surgery.* 2017;30(8):E1126-E1129.
58. Kim BJ, Kim ST, Jeong YG, Lee WH, Lee KS, Paeng SH. An efficient microvascular anastomosis training model based on chicken wings and simple instruments. *J Cerebrovasc Endovasc Neurosurg.* 2013;15(1):20-25.
59. Winkler-Schwartz A, Yilmaz R, Tran DH, et al. Creating a Comprehensive Research Platform for Surgical Technique and Operative Outcome in Primary Brain Tumor Neurosurgery. *World Neurosurgery.* 2020;144:e62-e71.
60. Chen P-C, Lin J-C, Chiang C-H, Chen Y-C, Chen J-E, Liu W-H. Engineering Additive Manufacturing and Molding Techniques to Create Lifelike Willis' Circle Simulators with Aneurysms for Training Neurosurgeons. *Polymers.* 2020;12(12).
61. Hu R, Barner KE, Steiner KV. A generalized haptic feedback approach for arbitrarily shaped objects. Paper presented at: MMVR2011.
62. Patel A, Koshy N, Ortega-Barnett J, et al. Neurosurgical tactile discrimination training with haptic-based virtual reality simulation. *Neurological Research.* 2014;36(12):1035-1039.
63. Peters TM, Linte CA, Yaniv Z, Williams J. Mixed and Augmented Reality in Medicine. In: Milton: Chapman and Hall/CRC; 2018.
64. Sugand K, Wescott RA, Carrington R, Hart A, van Duren BH. Training and Transfer Effect of FluoroSim, an Augmented Reality Fluoroscopic Simulator for Dynamic Hip Screw Guidewire Insertion: A Single-Blinded Randomized Controlled Trial. *J Bone Joint Surg Am.* 2019;101(17):e88.
65. Molina CA, Theodore N, Ahmed AK, et al. Augmented reality-assisted pedicle screw insertion: a cadaveric proof-of-concept study. *J Neurosurg Spine.* 2019:1-8.
66. Rojas-Muñoz E, Cabrera ME, Lin C, et al. The System for Telementoring with Augmented Reality (STAR): A head-mounted display to improve surgical coaching and confidence in remote areas. *Surgery.* 2020;167(4):724-731.
67. Kassab E, Tun JK, Arora S, et al. "Blowing up the barriers" in surgical training: exploring and validating the concept of distributed simulation. *Annals of surgery.* 2011;254(6):1059-1065.
68. Davids J, Manivannan S, Darzi A, Giannarou S, Ashrafian H, Marcus HJ. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. *Neurosurgical Review.* 2020.
69. Patel EA, Aydin A, Cearns M, Dasgupta P, Ahmed K. A Systematic Review of Simulation-Based Training in Neurosurgery, Part 1: Cranial Neurosurgery. *World Neurosurgery.* 2020;133:e850-e873.

70. Patel EA, Aydin A, Cearns M, Dasgupta P, Ahmed K. A Systematic Review of Simulation-Based Training in Neurosurgery, Part 2: Spinal and Pediatric Surgery, Neurointerventional Radiology, and Nontechnical Skills. *World neurosurgery.* 2020;133:e874-e892.

71. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. *International Journal of Computer Assisted Radiology and Surgery.* 2014;9(1):1-9.

72. Sabbagh AJ, Bajunaid KM, Alarifi N, et al. Roadmap for Developing Complex Virtual Reality Simulation Scenarios: Subpial Neurosurgical Tumor Resection Model. *World Neurosurgery.* 2020;139:e220-e229.

73. Bugdadi A, Sawaya R, Bajunaid K, et al. Is Virtual Reality Surgical Performance Influenced by Force Feedback Device Utilized? *Journal of Surgical Education.* 2019;76(1):262-273.

74. Bambakidis NC, Tomei KL. Editorial. Impact of COVID-19 on neurosurgery resident training and education. *Journal of Neurosurgery JNS.* 2020;133(1):10-11.

75. Zaed I, Tinterri B. Letter to the Editor: How is COVID-19 Going to Affect Education in Neurosurgery? A Step Toward a New Era of Educational Training. *World neurosurgery.* 2020;140:481-483.

76. Mirchi N, Ledwos N, Del Maestro RF. Intelligent Tutoring Systems: Re-Envisioning Surgical Education in Response to COVID-19. *Can J Neurol Sci.* 2020:1-3.

77. Clifton W, Damon A, Valero-Moreno F, Nottmeier E, Pichelmann M. The SpineBox: A Freely Available, Open-access, 3D-printed Simulator Design for Lumbar Pedicle Screw Placement. *Cureus.* 2020;12(4):e7738-e7738.

78. Abecassis IJ, Sen RD, Ellenbogen RG, Sekhar LN. Developing microsurgical milestones for psychomotor skills in neurological surgery residents as an adjunct to operative training: the home microsurgery laboratory. *J Neurosurg.* 2020:1-11.

79. Okland TS, Pepper J-P, Valdez TA. How do we teach surgical residents in the COVID-19 era? *Journal of Surgical Education.* 2020;77(5):1005-1007.

80. Fleiszer D, Hoover ML, Posel N, Razek T, Bergman S. Development and Validation of a Tool to Evaluate the Evolution of Clinical Reasoning in Trauma Using Virtual Patients. *Journal of surgical education.* 2018;75(3):779-786.

81. Sloth SB, Jensen RD, Seyer-Hansen M, Christensen MK, De Win G. Remote training in laparoscopy: a randomized trial comparing home-based self-regulated training to centralized instructor-regulated training. *Surgical Endoscopy.* 2021.

82. Davies P. Approaches to evidence-based teaching. *Medical Teacher.* 2000;22(1):14-21.

83. Chao TN, Frost AS, Brody RM, et al. Creation of an Interactive Virtual Surgical Rotation for Undergraduate Medical Education During the COVID-19 Pandemic. *Journal of Surgical Education.* 2021;78(1):346-350.

84. Schlich T. 'The Days of Brilliancy are Past': Skill, Styles and the Changing Rules of Surgical Performance, ca. 1820–1920. *Medical History.* 2015;59(3):379-403.

85. Carter BN. The fruition of Halsted's concept of surgical training. *Surgery.* 1952;32(3):518-527.

86. Harris KA, Nousiainen MT, Reznick R. Competency-based resident education-The Canadian perspective. *Surgery.* 2020;167(4):681-684.

87. Tobin S. Entrustable Professional Activities in Surgical Education. In: Nestel D, Dalrymple K, Paige JT, Aggarwal R, eds. *Advancing Surgical Education: Theory, Evidence and Practice.* Singapore: Springer Singapore; 2019:229-238.

88. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher.* 2013;35(7):564-568.

89. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *BJS (British Journal of Surgery).* 1997;84(2):273-278.

90. MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued Validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the Simulated Environment. *Teaching and Learning in Medicine.* 2016;28(1):72-79.

91. Fecso AB, Szasz P, Kerezov G, Grantcharov TP. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. *Annals of Surgery.* 2017;265(3):492-501.

92. Ilgen JS, Ma IWY, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical Education.* 2015;49(2):161-173.

93. de Montbrun S, Satterthwaite L, Grantcharov TP. Setting pass scores for assessment of technical performance by surgical trainees. *Br J Surg.* 2016;103(3):300-306.

94. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Advances in Health Sciences Education.* 2015;20(5):1149-1175.

95. Hadley C, Lam SK, Briceño V, Luerssen TG, Jea A. Use of a formal assessment instrument for evaluation of resident operative skills in pediatric neurosurgery. *Journal of Neurosurgery: Pediatrics PED.* 2015;16(5):497-504.

96. Winkler-Schwartz A, Marwa I, Bajunaid K, et al. A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study. *World Neurosurgery.* 2019;127:e230-e235.

97. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med.* 2019;25(1):30-36.

98. Levin M, McKechnie T, Kruse CC, Aldrich K, Grantcharov TP, Langerman A. Surgical data recording in the operating room: a systematic review of modalities and metrics. *British Journal of Surgery.* 2021.

99. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F. Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance. *JAMA Network Open.* 2020;3(3):e201664-e201664.

100. Jin A, Yeung S, Jopling J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. Paper presented at: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)2018.

101. Liu D, Jiang T. Deep Reinforcement Learning for Surgical Gesture Segmentation and Classification. Paper presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018; 2018//, 2018; Cham.

102. Binkley CE, Green BP. Does Intraoperative Artificial Intelligence Decision Support Pose Ethical Issues? *JAMA Surgery.* 2021.

103. Eysenbach G, Ho W-H, Kolachalama V, Chan KS, Zary N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. *JMIR Medical Education.* 2019;5(1).

104. Flores CD, Barros P, Cazella S, Bez MR. Leveraging the learning process in health through clinical cases simulator. Paper presented at: 2013 IEEE 2nd International Conference on Serious Games and Applications for Health (SeGAH); 2-3 May 2013, 2013.

105. Hamdy H, Al-Moslih A, Tavarnesi G, Laus A. Virtual patients in problem-based learning. *Medical Education.* 2017;51(5):557-558.

106. McFadden P, Crim A. Comparison of the Effectiveness of Interactive Didactic Lecture Versus Online Simulation-Based CME Programs Directed at Improving the Diagnostic Capabilities of Primary Care Practitioners. *Journal of Continuing Education in the Health Professions.* 2016;36(1).

107. Sheikh AY, Fann JI. Artificial Intelligence: Can Information be Transformed into Intelligence in Surgical Education? *Thorac Surg Clin.* 2019;29(3):339-350.

108. Van Melle E, Frank JR, Holmboe ES, Dagnone D, Stockley D, Sherbino J. A Core Components Framework for Evaluating Implementation of Competency-Based Medical Education Programs. *Acad Med.* 2019;94(7):1002-1009.

109. Eppich W, Cheng A. Promoting Excellence and Reflective Learning in Simulation (PEARLS): Development and Rationale for a Blended Approach to Health Care Simulation Debriefing. *Simulation in Healthcare.* 2015;10(2).

110. Pane JF, Griffin BA, McCaffrey DF, Karam R. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis.* 2014;36(2):127-144.

111. Harley JM, Bouchet F, Azevedo R. Aligning and Comparing Data on Emotions Experienced during Learning with MetaTutor. Paper presented at: Artificial Intelligence in Education; 2013//, 2013; Berlin, Heidelberg.

112. Bouchet F, Harley J, M, Azevedo R. Evaluating Adaptive Pedagogical Agents' Prompting Strategies Effect on Students' Emotions. Paper presented at: 14th International Conference on Intelligent Tutoring Systems (ITS 2018); 2018-06-11, 2018; Montreal, Canada.

113. Sweller J. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review.* 2010;22(2):123-138.

114. Fraser K, Ma I, Teteris E, Baxter H, Wright B, McLaughlin K. Emotion, cognitive load and learning outcomes during simulation training. *Med Educ.* 2012;46(11):1055-1062.

115. Lawrence C. Medical Minds, Surgical Bodies. In: Lawrence C, Shapin S, eds. *Science Incarnate: Historical Embodiments of Natural Knowledge.* University of Chicago Press; 1998:156--201.

116. Reznick R, Harris, K., Horsely, T., Sheikh Hassani, M. *Task Force Report on Artificial Intelligence and Emerging Digital Technologies.* The Royal College of Physicians and Surgeons of Canada; February 2020.

117. Munro C, Burke J, Allum W, Mortensen N. Covid-19 leaves surgical training in crisis. *BMJ.* 2021;372:n659.

118. Samuel BT, Benjamin KH, Aaron AC-G. Editorial. Innovations in neurosurgical education during the COVID-19 pandemic: is it time to reexamine our neurosurgical training models? *Journal of Neurosurgery JNS.* 2020;133(1):14-15.

119. A. Butt K, Augestad KM. Educational value of surgical telementoring. *Journal of Surgical Oncology.* 2021;124(2):231-240.

120. Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *The Lancet Digital Health.* 2020;2(10):e537-e548.

121. Lynch J, Aughwane P, Hammond TM. Video Games and Surgical Ability: A Literature Review. *Journal of Surgical Education.* 2010;67(3):184-189.

122. Rui M, Lee JE, Vauthey J-N, Conrad C. Enhancing surgical performance by adopting expert musicians' practice and performance strategies. *Surgery.* 2018;163(4):894-900.

123. Macnamara BN, Moreau D, Hambrick DZ. The relationship between deliberate practice and performance in sports: A meta-analysis. *Perspectives on Psychological Science.* 2016;11(3):333-350.

124. McGaghie WC. Mastery Learning: It Is Time for Medical Education to Join the 21st Century. *Academic Medicine.* 2015;90(11).

125. Kolb DA. *Experiential learning: Experience as the source of learning and development.* FT press; 2014.

126. Dean WH, Gichuhi S, Buchan JC, et al. Intense Simulation-Based Surgical Education for Manual Small-Incision Cataract Surgery: The Ophthalmic Learning and Improvement Initiative in Cataract Surgery Randomized Clinical Trial in Kenya, Tanzania, Uganda, and Zimbabwe. *JAMA Ophthalmology.* 2021;139(1):9-15.

127. Meling TR, Meling TR. The impact of surgical simulation on patient outcomes: a systematic review and meta-analysis. *Neurosurgical Review.* 2021;44(2):843-854.

128. Lohre R, Bois AJ, Athwal GS, Goel DP, on behalf of the Canadian S, Elbow Society. Improved Complex Skill Acquisition by Immersive Virtual Reality Training: A Randomized Controlled Trial. *JBJS.* 2020;102(6).

129. Barry Issenberg S, McGaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical Teacher.* 2005;27(1):10-28.

130. Schaffir J, Strafford K, Worly B, Traugott A. Challenges to Medical Education on Surgical Services During the COVID-19 Pandemic. *Medical Science Educator.* 2020;30(4):1667-1671.

131. Eppich WJ, Hunt EA, Duval-Arnould JM, Siddall VJ, Cheng A. Structuring Feedback and Debriefing to Achieve Mastery Learning Goals. *Academic Medicine.* 2015;90(11):1501-1508.

132. Janzen KJ, Jeske S, MacLean H, et al. Handling Strong Emotions Before, During, and After Simulated Clinical Experiences. *Clinical Simulation in Nursing.* 2016;12(2):37-43.

133. Bilgic E, Turkdogan S, Watanabe Y, et al. Effectiveness of Telementoring in Surgery Compared With On-site Mentoring: A Systematic Review. *Surg Innov.* 2017;24(4):379-385.

134. Erridge S, Yeung DKT, Patel HRH, Purkayastha S. Telementoring of Surgeons: A Systematic Review. *Surg Innov.* 2019;26(1):95-111.

135. Haji FA. Simulation in Neurosurgical Education During the COVID-19 Pandemic and Beyond. *Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques.* 2021;48(2):152-154.

136. Raghu Prasad MS, Manivannan M. Comparison of Force Matching Performance in Conventional and Laparoscopic Force-Based Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting.* 2014;58(1):683-687.

137. Lee SH, Jin SH, An J. The difference in cortical activation pattern for complex motor skills: A functional near- infrared spectroscopy study. *Scientific Reports.* 2019;9(1):14066.

138. Bravi R, Cohen EJ, Martinelli A, Gottard A, Minciacchi D. When Non-Dominant Is Better than Dominant: Kinesiotape Modulates Asymmetries in Timed Performance during a Synchronization-Continuation Task. *Front Integr Neurosci.* 2017;11:21-21.

139. Adeleye AO, Fasunla JA, Young PH. Skull base surgery in a large, resource-poor, developing country with few neurosurgeons: prospects, challenges, and needs. *World neurosurgery.* 2012;78(1-2):35-43.

140. Ladd BM, Tackla RD, Gupte A, et al. Feasibility of Telementoring for Microneurosurgical Procedures Using a Microscope: A Proof-of-Concept Study. *World Neurosurgery.* 2017;99:680-686.

141. Malekzadeh M, Mustafa MB, Lahsasna A. A review of emotion regulation in intelligent tutoring systems. *Journal of Educational Technology & Society.* 2015;18(4):435-445.

142. D'mello S, Graesser A. AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans Interact Intell Syst.* 2013;2(4):Article 23.

143. Grawemeyer B, Mavrikis M, Holmes W, Gutiérrez-Santos S, Wiedmann M, Rummel N. Affective learning: improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction : The Journal of Personalization Research.* 2017;27(1):119-158.

144. Edmondson AC. The competitive imperative of learning. *Harv Bus Rev.* 2008;86(7-8):60-67, 160.

TABLES

Table 1. OSATS Scale Consistency and Inter-Rater Reliability

OSATS Category	Intraclass Correlation Coefficient	Cronbach's Alpha	Pearson's Correlation Coefficient	Mean Score Rater 1 (± SD)	Mean Score Rater 2 (± SD)
Instrument Handling	0.913	0.908	0.839	4.67 (± 1.21)	4.50 (± 1.37)
Respect for Tissue	0.845	0.834	0.738	4.50 (± 2.43)	4.00 (± 1.90)
Hemostasis	0.819	0.821	0.702	4.17 (± 2.64)	3.33 (± 2.34)
Economy of Movement	0.828	0.802	0.825	3.83 (± 1.17)	4.00 (± 2.28)
Flow	0.786	0.755	0.647	4.00 (± 1.67)	3.83 (± 2.40)
Overall	0.836	0.809	0.717	3.83 (± 1.72)	3.83 (± 2.40)

Table 1. OSATS Scale Consistency and Inter-Rater Reliability. Instructors' assessment results of pre-recorded medical student performance at the end of the standardized instructor training. Both instructors rated the same 20 pre-recorded videos of medical students' performance. Intraclass correlation coefficient (ICC) is a measure of the inter-rater reliability. ICC values between 0.75-0.90 indicate good reliability. Cronbach's alpha (α) is a measure of how closely related a set of scale items are. α values between 0.70-0.90 indicate good internal consistency. Positive correlation between both raters on. Individual items of the OSATS scale.

Table 2. Demographic Characteristics of Included Participants.

	Medical Student, No. (%)		
Characteristic	**Control Group (n = 23)**	**VOA Group (n = 23)**	**Instructor Group (n = 24)**
Age, mean (SD)	21.7 (2.4)	21.9 (2.5)	21.8 (2.1)
Gender			
Male	9 (39)	10 (43)	10 (42)
Female	14 (61)	13 (57)	14 (58)
Undergraduate Medical Training Level			
Med-P (Preparatory)[a]	9 (39)	10 (43)	7 (29)
First Year	8 (35)	8 (35)	9 (38)
Second Year	6 (26)	5 (22)	8 (33)
Institution			
McGill University	14 (61)	8 (35)	10 (42)
University of Montreal	3 (13)	7 (30)	7 (29)
University of Laval	6 (26)	7 (30)	6 (25)
University of Sherbrooke	0	1 (5)	1 (4)
Dominant hand			
Right	23 (100)	21 (91)	22 (92)
Left	0	2 (9)	2 (8)
Interest in pursuing surgery, mean (SD)[b]	3.7 (1.0)	3.9 (1.1)	3.8 (1.2)
Video Games			
Not at all	15 (65)	15 (65)	16 (67)

	Medical Student, No. (%)		
Characteristic	**Control Group (n = 23)**	**VOA Group (n = 23)**	**Instructor Group (n = 24)**
1-5 hrs/wk	5 (22)	6 (26)	5 (21)
6-10 hrs/wk	2 (9)	2 (9)	2 (8)
>11 hrs/wk	1 (4)	0	1 (4)
Play musical instruments			
Yes	12 (52)	8 (35)	13 (54)
No	11 (48)	15 (65)	11 (46)
Did competitive sports in the past 5 years			
Yes	12 (52)	17 (74)	17 (71)
No	11 (48)	6 (26)	7 (29)
Prior VR experience in any domain			
None	14 (61)	12 (52)	12 (50)
Passive (Google Earth, Videos, etc.)	8 (35)	10 (43)	9 (38)
Active (Games, Simulation, etc.)	1 (4)	1 (5)	3 (12)
Prior experience with any VR surgical simulator			
Yes	1 (4)	0	0
No	22 (96)	23 (100)	24 (100)

Abbreviations: VOA, Virtual Operative Assistant; Med-P, Medicine Preparatory; VR, Virtual Reality

[a] Medicine Preparatory (Med-P) is a one-year preparatory program for graduates of the Quebec Collegial (CEGEP) system who have been offered a position from the medical program of McGill University or University of Montreal.

[b] Rated on a Likert Scale (1-5), with 1 indicating less interest and 5 indicating more interest.

Table 3. Intervention Comparison Table

	Virtual Operative Assistant	**Remote Expert Instruction**
Task Goal	Complete resection of the tumor with minimal bleeding and damage to healthy tissues.	
Learning Objectives (Competency Criteria)	**Safety:** 1. Maximum force applied with the bipolar forceps 2. Mean rate of bleeding **Movement:** 3. Mean instrument tip separation 4. Mean Acceleration of the bipolar forceps	**Safety:** 1. Respect for Tissue 2. Hemostasis **Movement:** 3. Instrument Handling 4. Economy of Movement 5. Flow **Overall Quality:** 6. Overall
Performance Assessment Tool	Criteria-based assessment using a machine learning classifier algorithm. Four AI-selected metrics used by a support vector machine for performance classification and quantitative benchmark evaluation.	Criteria-based assessment using the Video Assessment Sheet (Appendix 2). Six relevant performance categories selected by experts for performance assessment on a 7-point Likert scale.
Learning Theory	Mastery learning through deliberate practice guided by self-regulated learning.	Mastery learning through deliberate practice guided by self-regulated learning.
Feedback Delivery	Audiovisual metric-specific feedback provided autonomously and immediately depending on the participant's competency.	Live verbal debriefing with scripted feedback and instructions provided immediately depending on the participant's competency.
Feedback Content	Metric-specific videos played based on the learner's individual needs that describe the appropriate assessment criteria, demonstrate novice and expert performance examples, and provide actionable instructions to excel. Senior consultants with extensive subpial experience provided instructions and performance in the videos.	OSATS category-specific feedback prompts and actionable instructions used in a debriefing script that describes the relevant performance category and the lacking competency, and provides instructions tailored to the learner's individual needs. Feedback prompts and instructions were provided by senior consultants on how to excel.

FIGURES

Figure 1. The NeuroVR Simulator Platform

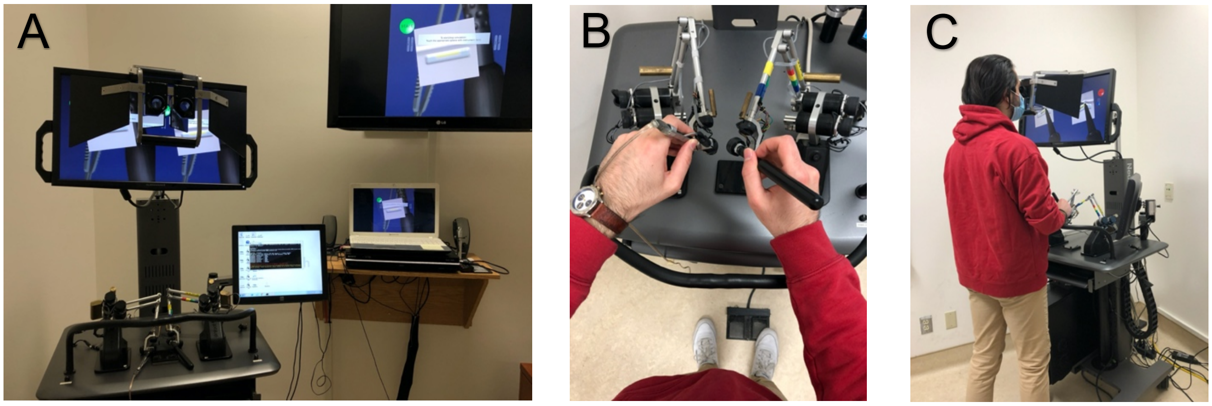

Figure 1. The NeuroVR Simulator Platform. (A) The NeuroVR simulator with the practice subpial scenario on the screen. **(B)** Participant using the handles for subpial tumor resection (bipolar instrument with the left hand and the aspirator with the right hand) foot pedals (at the bottom of the image) control the activation of the corresponding instruments. **(C)** Participant viewing the screen through the stereoscope and performing the practice subpial scenario.

Figure 2. Practice Subpial Tumor Resection Scenario

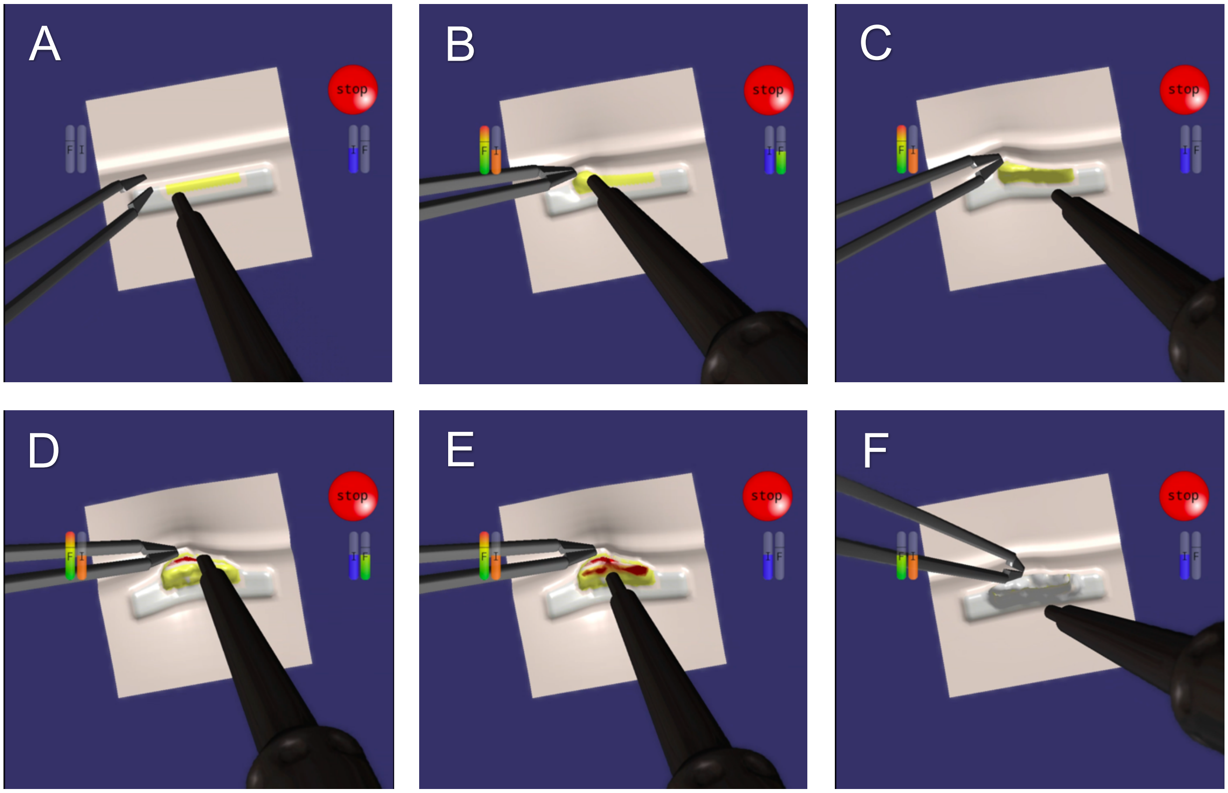

Figure 2. Practice Subpial Tumor Resection Scenario. (A) Practice subpial scenario at the start of the simulation, yellow tissue represents the tumor, instrument on the left is the bipolar and the instrument on the right is the aspirator. **(B)** Participant using the bipolar to lift the pia and aspirator utilized to resect the tumor lying beneath the pia. **(C)** Appearance following resection of superficial tumor. Yellow tissue remaining depicts the deeper tumor areas. **(D)** Participant exposing the simulated deep cerebral vessel (red). **(E)** Instrument injury to the blood vessel resulting in bleeding. **(F)** Complete resection of the tumor.

Figure 3. Realistic Subpial Tumor Resection Scenario

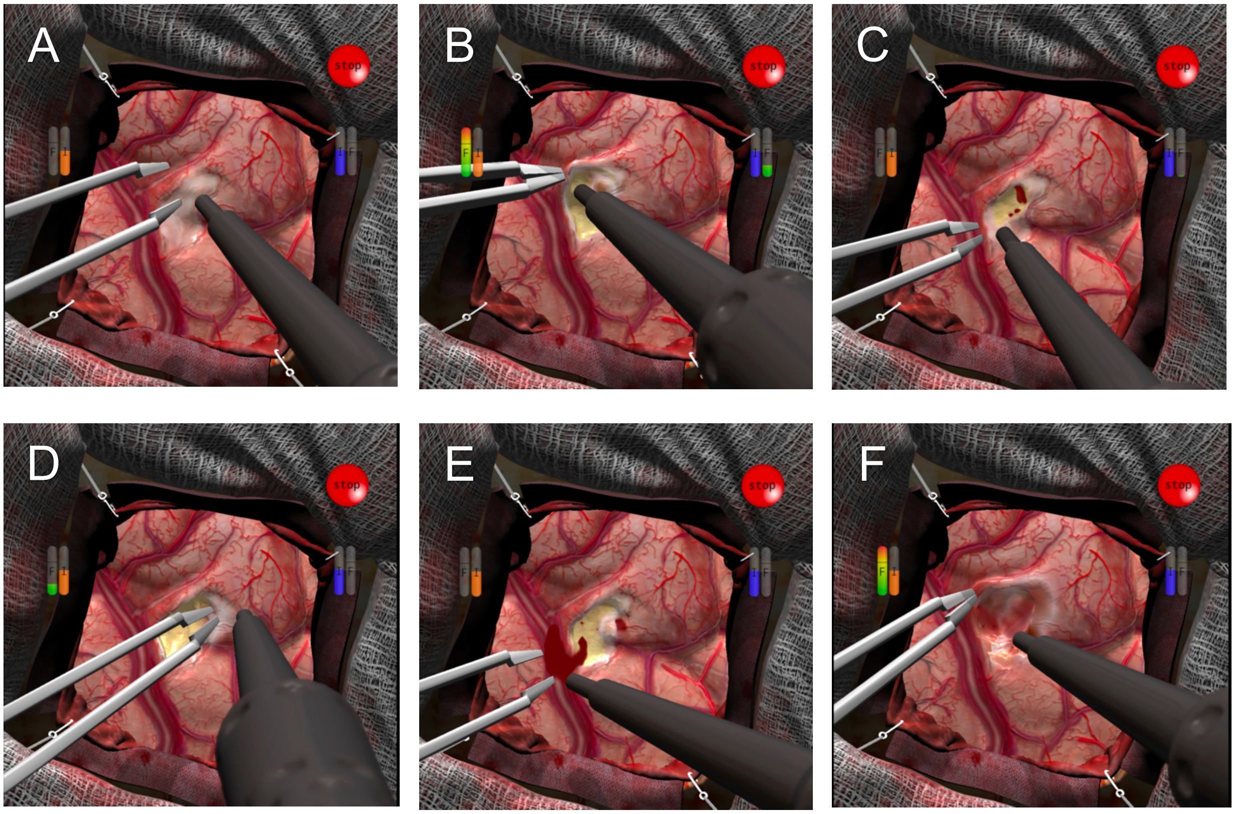

Figure 3. Realistic Subpial Tumor Resection Scenario. (A) Realistic subpial scenario at the start of the simulation, off-white tissue represents the tumor. **(C)** Participant while using the aspirator causes minor bleeding from the tumor. **(D)** Participant cauterizing bleeding points. **(E)** Injury to the superficial cerebral vein followed by significant bleeding. **(B)** Participant using the simulated bipolar to lift pia and begins resecting tumor with the simulated aspirator. **(F)** Completed Tumor resection.

Figure 4. CONSORT Flow Diagram

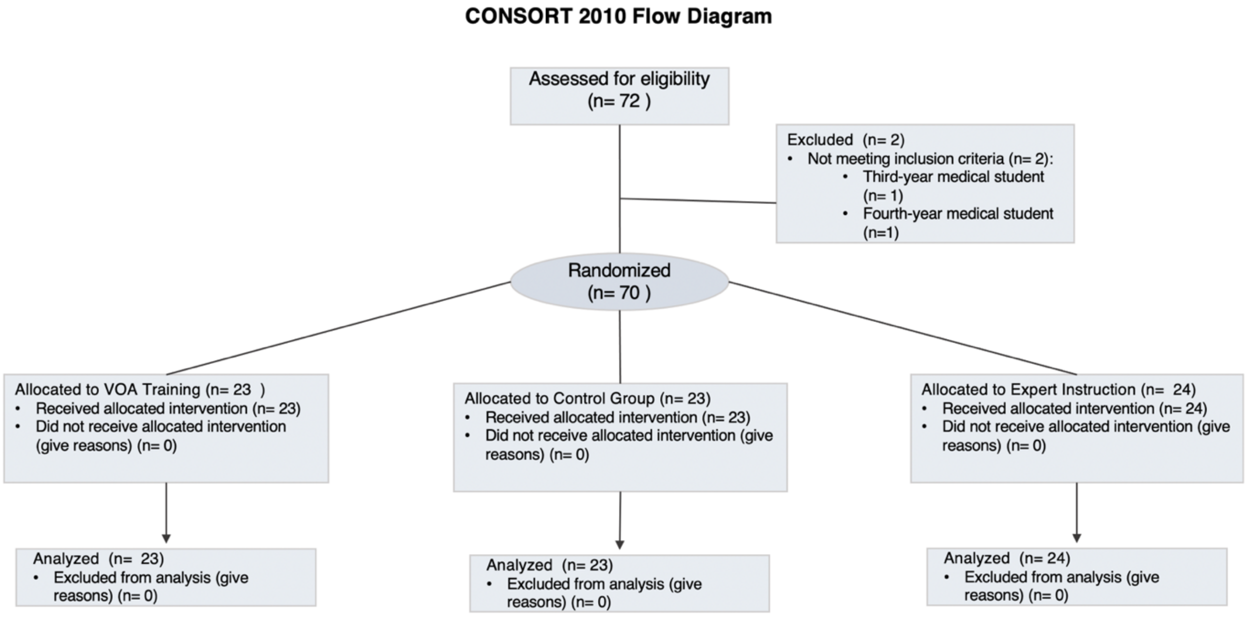
CONSORT 2010 Flow Diagram
Assessed for eligibility (n= 72)
Excluded (n= 2)
• Not meeting inclusion criteria (n= 2):
• Third-year medical student (n= 1)
• Fourth-year medical student (n=1)
Randomized (n= 70)
Allocated to VOA Training (n= 23)
• Received allocated intervention (n= 23)
• Did not receive allocated intervention (give reasons) (n= 0)
Allocated to Control Group (n= 23)
• Received allocated intervention (n= 23)
• Did not receive allocated intervention (give reasons) (n= 0)
Allocated to Expert Instruction (n= 24)
• Received allocated intervention (n= 24)
• Did not receive allocated intervention (give reasons) (n= 0)
Analyzed (n= 23)
• Excluded from analysis (give reasons) (n= 0)
Analyzed (n= 23)
• Excluded from analysis (give reasons) (n= 0)
Analyzed (n= 24)
• Excluded from analysis (give reasons) (n= 0)

Figure 4. CONSORT Flow Diagram. 72 participants were assessed for eligibility. Two were excluded for not meeting the inclusion criteria of being enrolled in Medicine Preparatory, first, or second year of medical school. 70 participants were stratified by gender and block randomized to one of three arms. Data from all included participants was used for the final intention-to-treat analysis.

Figure 5. Learning with the Virtual Operative Assistant (VOA)

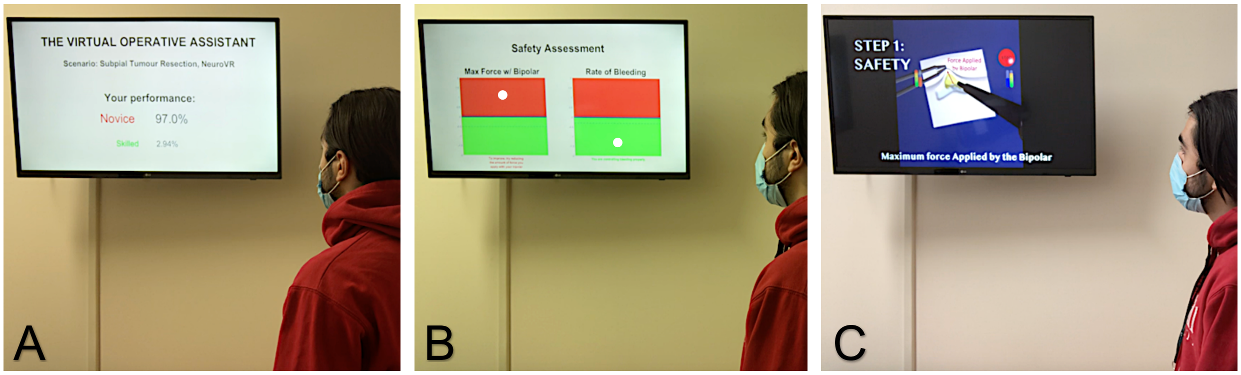

Figure 5. Learning with the Virtual Operative Assistant (VOA). (A) Participant viewing the VOA's performance prediction of their practice subpial resection. **(B)** Participant viewing a breakdown of their performance assessment on two safety metrics. A score in the red box (depicted by the white dot) suggests falling outside the competence benchmark for that metric. **(C)** VOA plays the appropriate feedback video for the metric that needs improvement.

Figure 6. Remote Expert Instruction

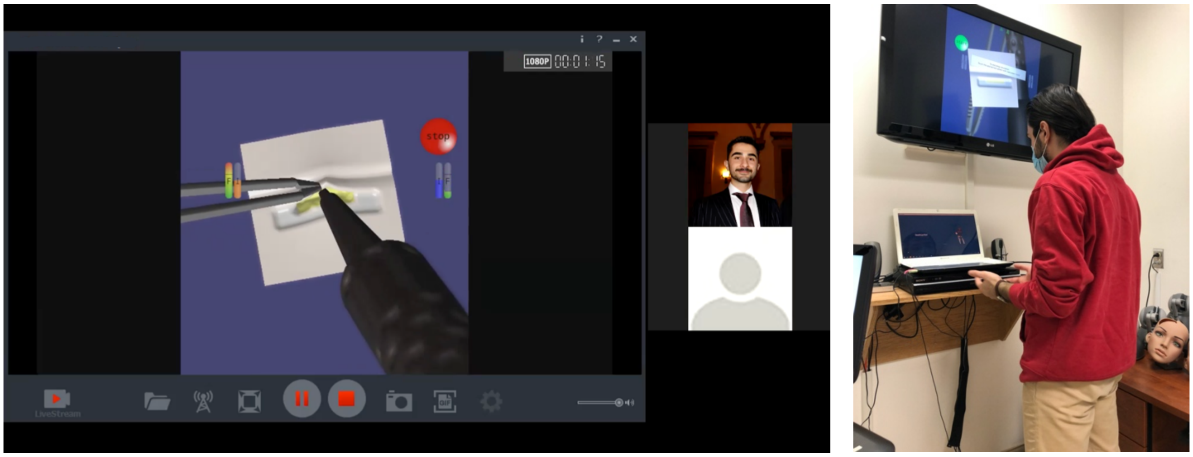

Figure 6. Remote Expert Instruction. Left. Livestream on-screen performance of a participant’s practice resection shared virtually with an instructor located remotely. **Right.** Participant debriefing and receiving feedback from the remote instructor after the simulation resection through a videotelephony software.

Figure 7. Participant's Progression Through the VOA Training.

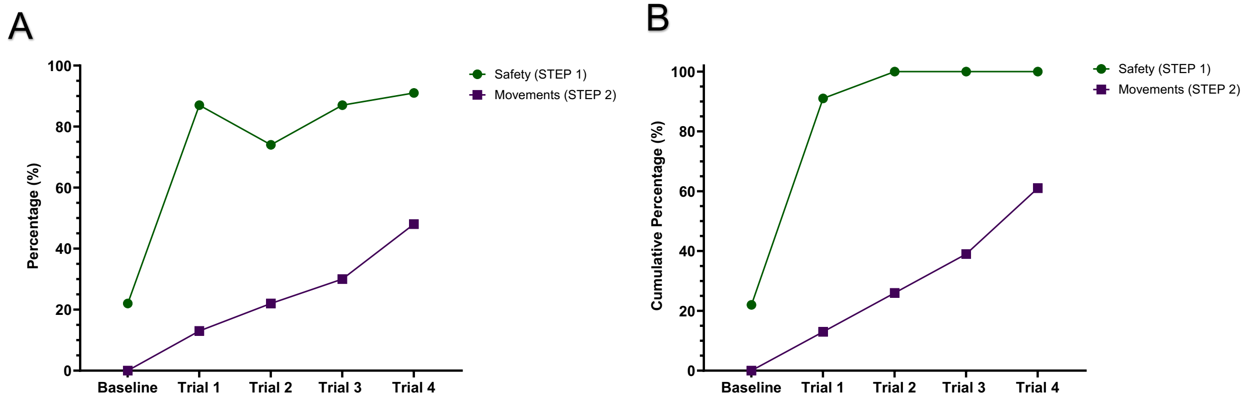

Figure 7. Participant's Progression Through the VOA Training. (A) Percentage of VOA participants who passed STEP-1 (safety) and STEP-2 (instrument movements) of VOA training at a specific trial. **(B)** Cumulative percentage of VOA participants who passed a specific VOA competency on or before a given trial. Data shows that the proportion of individuals who passed a competency at a given trial, were likely to pass that competency again in the following trial.

Figure 8. Performance Assessment in the Practice Tumor Resections.

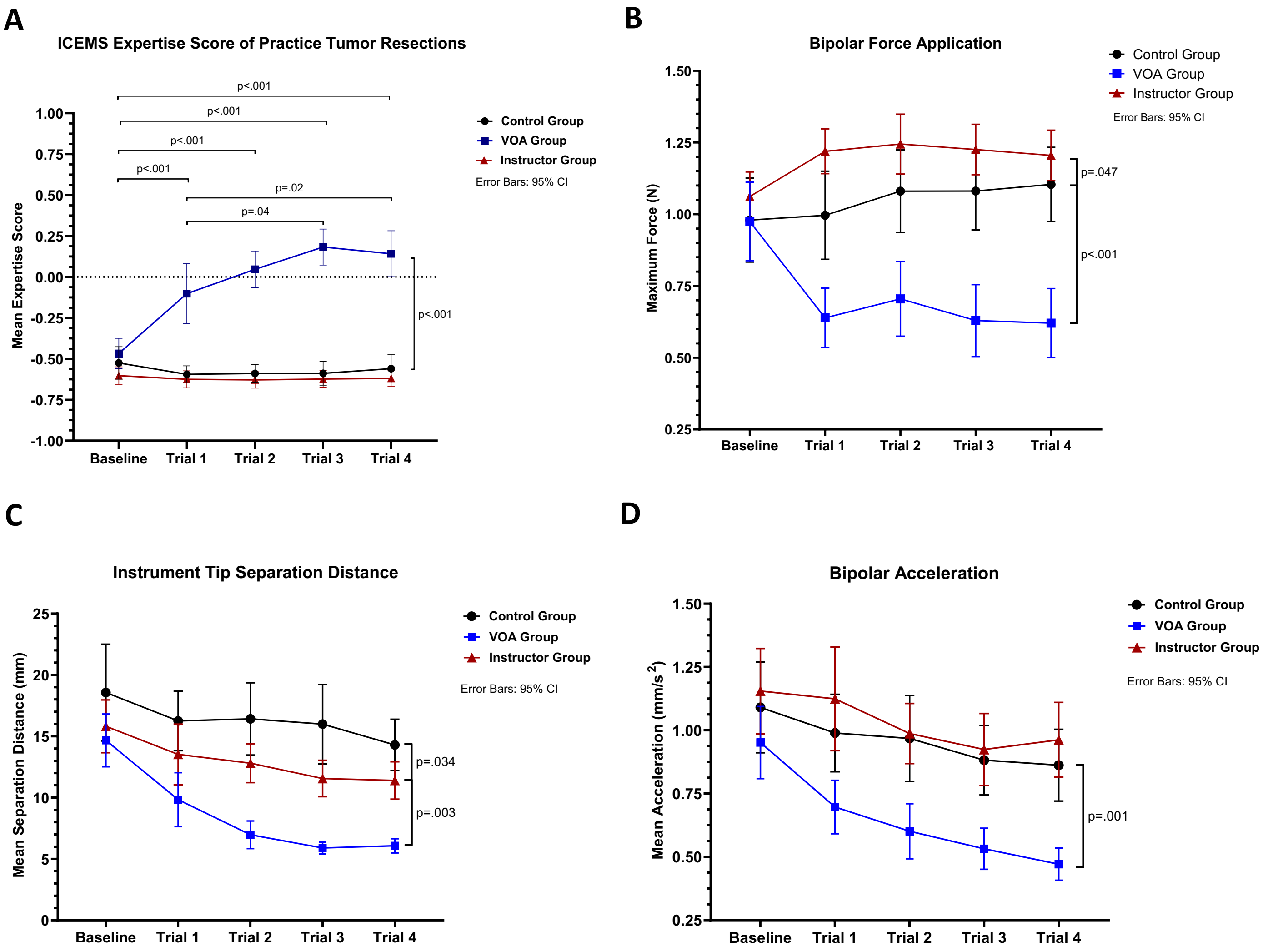

Figure 8. Performance Assessment in the Practice Tumor Resection Simulations. (A) Mean ICEMS Expertise Scores. Negative scores correspond to a novice and a positive score corresponds to a more expert performance. Scores in each trial are the average of all predictions made for every 200-milliseconds of the simulated procedure (about 1500 predictions for a 5-minute practice scenario). (**B)** Maximum bipolar force application is a recording of the highest amount of force applied with the bipolar during the entire operation. (**C)** Average instrument tip separation distance measured as the mean distance between the aspirator and the bipolar tips.

(D) Average bipolar acceleration measured as the rate of change in the bipolar instrument's velocity. All error bars represent the 95% Confidence Interval of the mean and p-values were adjusted by Bonferroni correction for multiple tests. Two-way mixed ANOVA with trial number as the within-subjects variable, type of feedback as the between-subjects variable, and baseline performance as a covariate, was conducted for the Expertise Scores and metrics data. P-values from Tukey's HSD multiple comparisons post-hoc test is reported for the between-group differences. One-way repeated measures ANOVA was conducted to assess pairwise differences of VOA group's Expertise Scores between trials.

Figure 9. Performance Assessment in the Realistic Tumor Resection.

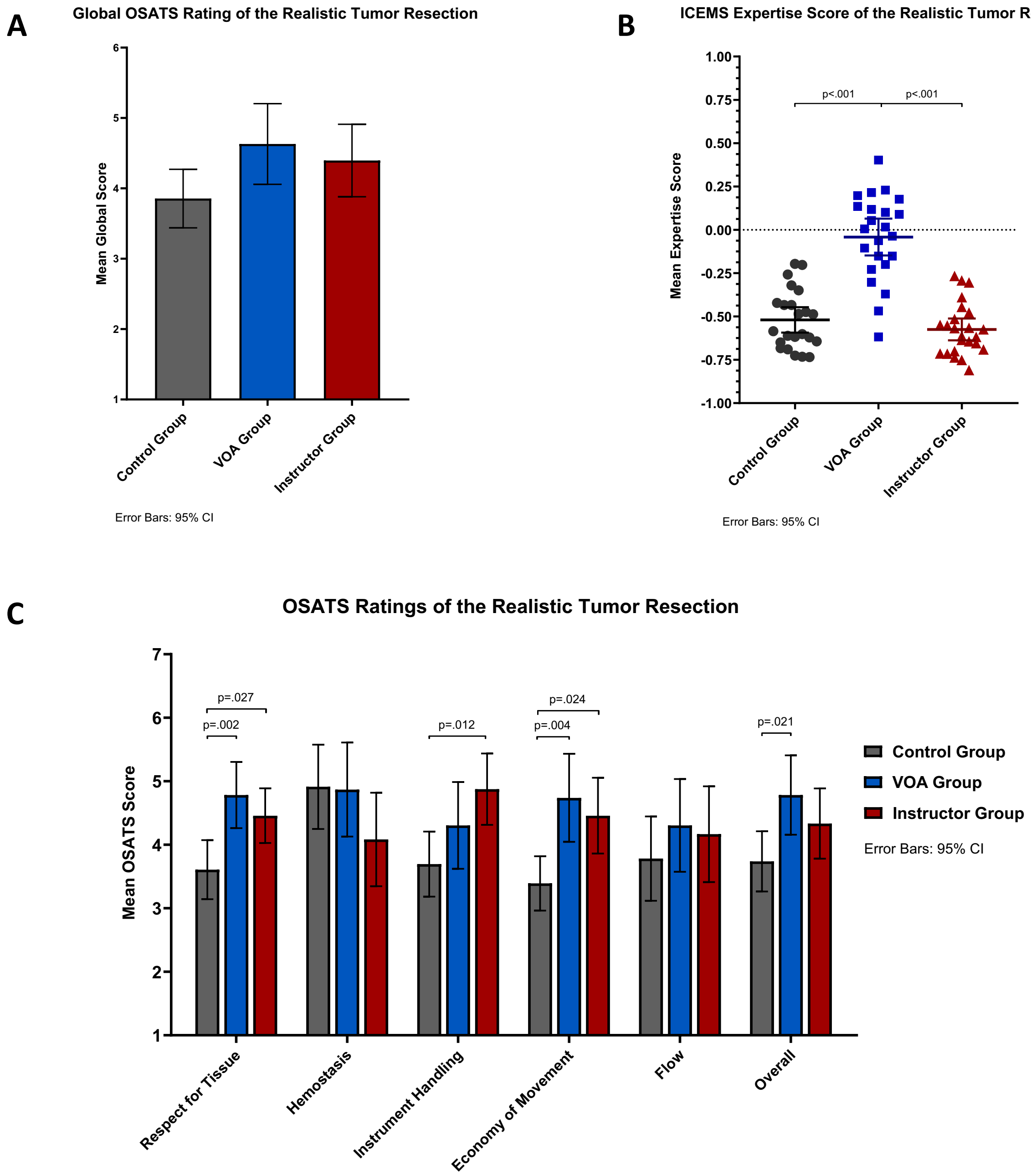

Figure 9. Performance Assessment in the Realistic Tumor Resection Simulation. (A) Mean ICEMS Expertise Score. One-way ANOVA found statistical difference between three groups (p<.001), Tukey's HSD post-hoc analysis identified difference between the VOA condition and both the control (p<.001) and instructor (p<.001) groups. (**B)** Mean global OSATS rating of the

three groups. The global rating is the average score of all 6 items of the OSATS scale. One-way ANOVA found no statistical difference between groups (p=.081). (**C)** Mean performance score for individual items of the OSATS scale for three groups. One-way ANOVA identified a statistical difference in respect for tissue (p=.002), instrument handling (p=.017), economy of movement (p=.003) and overall (p=.028). No significant difference was found in hemostasis (p=.164) and flow (p=.552). Tukey's post-hoc analysis of significant findings with Bonferroni correction of multiple tests identified significant difference between the VOA and control group in overall (p=.021), a significant difference between the instructor and control group in instrument handling (p=.012), and that both the VOA and instructor groups had significantly higher respect for respect for tissue (p=.002, VOA; p=.027, instructor) and economy of movement (p=.004, VOA; p=.024, instructor) compared to control.

Figure 10. Emotions and Cognitive Load Throughout the Simulation Training

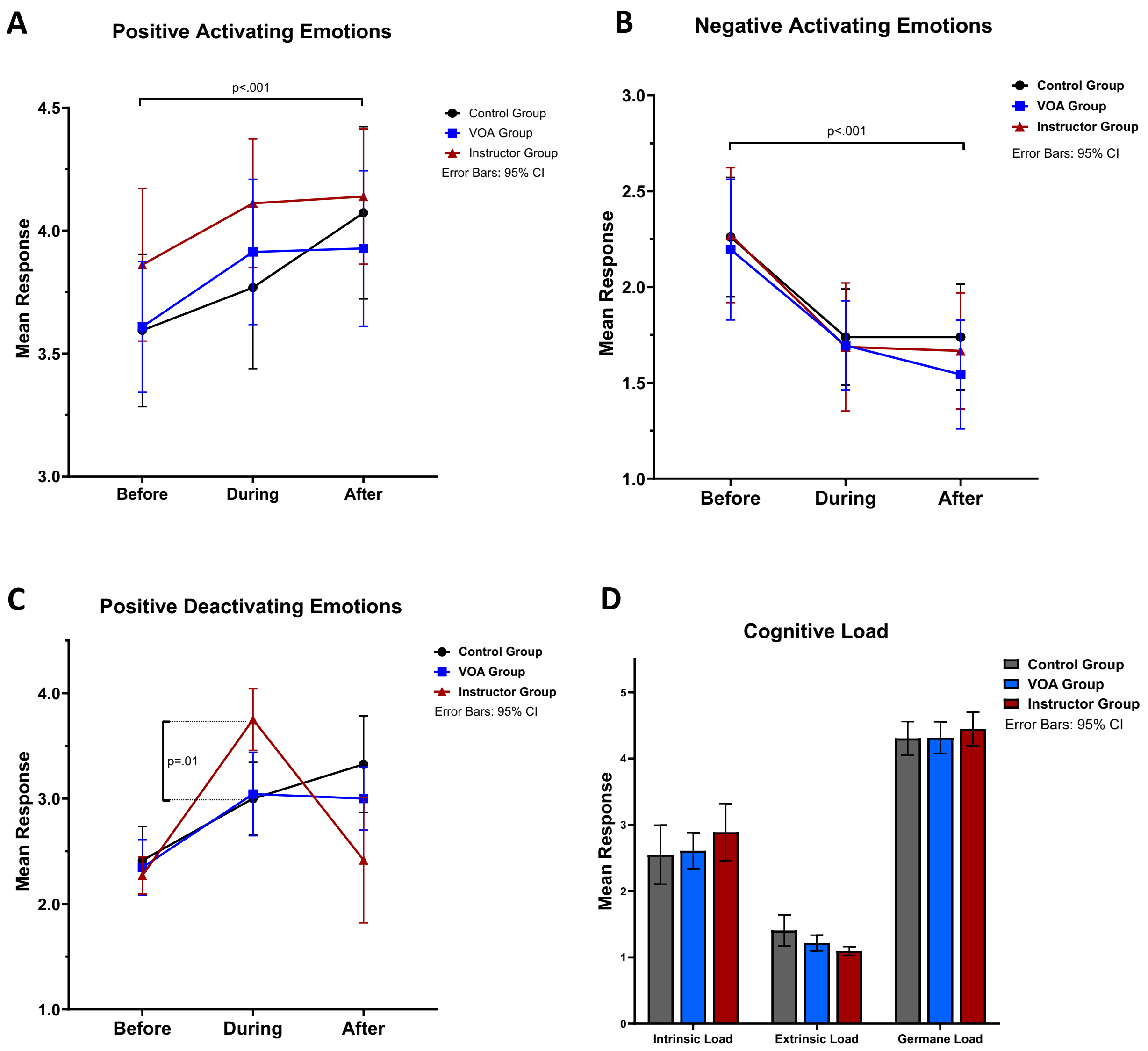

Figure 4. Emotions and Cognitive Load Throughout the Simulation Training. A) Positive Activating emotions (happy, hopeful, grateful) and **B)** Negative Activating emotions (confusion, anxiety) showed a significant within-subjects effect (p<.001) with no significant between-subjects difference. **C)** Positive Deactivating emotions (relaxed, relieved) show a group*time interaction effect for participants in the Instructor Group during the training. Tukey's HSD post-

hoc test showed a significant difference between the instructor group and both VOA (p=.010) and control (p=.006) groups in the strength of positive deactivating emotions during training. **D)** Cognitive Load Index (CLI) responses. One-way ANOVA did not find statistical differences between groups in intrinsic, extrinsic, and germane cognitive load. Error bars represent the 95% Confidence Interval of the mean.

Figure 11. Instructor Group's OSATS Ratings During Practice

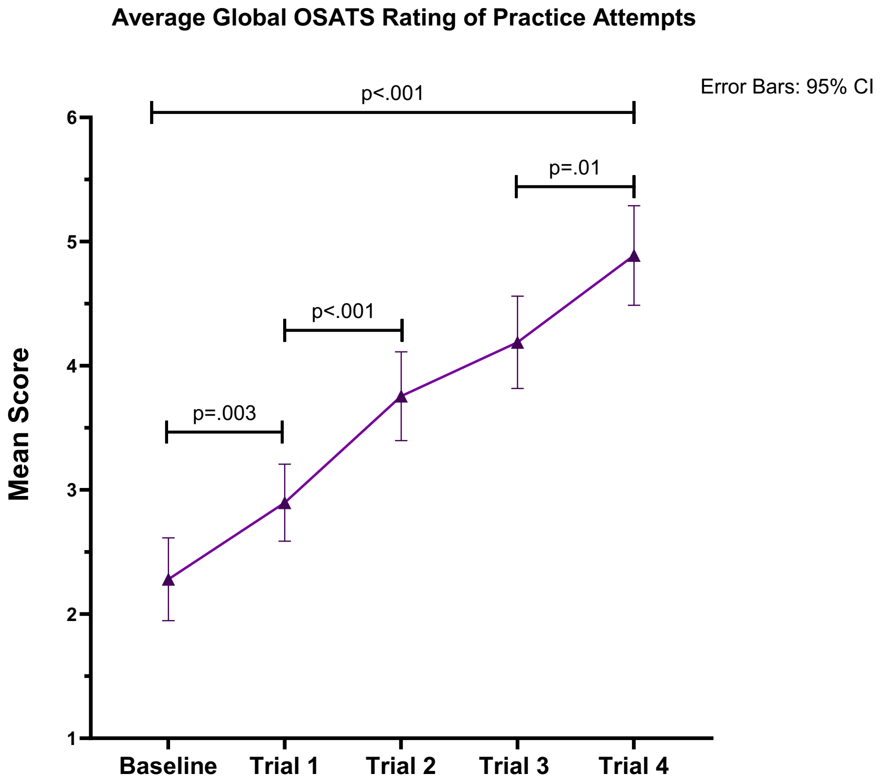

Figure 11. Instructor Group's OSATS Ratings During Practice. Average global OSATS ratings during practice scenario, measured as the mean of the 6 items in the visual rating scale. Instructor group's global OSATS ratings during practice shows performance improvement after debriefing and feedback sessions. In this group, average global OSATS scores improved by 0.62 points (95% CI 0.17-1.07, p=.003) from baseline at Trial-1, 0.86 points (95% CI 0.45-1.27, p<.001) from Trial-1 to Trial-2, 0.44 points (95% CI -0.09-0.95, p=.17) from Trial-2 to Trial-3, and 0.70 points (95% CI 0.14-1.26, p=.01) from Trial-3 to Trial-4. Figure above depicts the learning curve for participants in the instructor Group based on the OSATS scale. Bars represent the 95% confidence interval of the mean.

APPENDIX

Appendix 1. Standardized Instructor Training Protocol

Objective:

To adapt the traditional apprenticeship learning model to a remote context, we need to ensure that the study instructors are trained to:

1. Perform the simulated practice and realistic subpial resections expertly
2. Rate students' performance from screen-recorded videos with consistency and reliability
3. Deliver constructive feedback in scripted debriefing sessions.

Methods:

Eight 90-minute learning sessions in a two-week workshop, provided two senior neurosurgery resident instructors (A.A., M.B., Post Graduate Year 5) with standardized training to become proficient at leading virtual pedagogical sessions remotely for medical student participants of this study.

Two sessions involved performing the simulated resections under the supervision of a senior consultant, who demonstrated the technical competencies required, explained OSATS's qualitative assessment criteria, and described how to lead an effective debriefing based on the PEARLS model.

In the following sessions, instructors trained independently through deliberate practice guided by self-regulated learning where they graded their own screen-recorded performance using the Assessment Sheet (Appendix 2).

Because in traditional apprenticeship experts in the operating room have no access to performance metrics and assessment depends only on visual rating, instructors were blinded and unaware of the AI assessment metrics to best replicate the current intraoperative instruction and reduce potential bias in their assessment and instruction in the study.

At the end of the training, instructors were evaluated by the senior consultant based on their ability to achieve technical competence in both simulation resections and lead scripted debriefing sessions remotely. Scale consistency and inter-rater reliability was determined from instructor ratings of 20 randomly selected videos of medical students' performance of both realistic and practice subpial simulations.

Theories used in training:

This training utilized two key educational theories: Deliberate practice and self-regulated learning (SRL). Both deliberate practice and SRL accelerated learning by leveraging effective learning strategies such as drawing upon reflective observation through self-assessment and using forethought to set specific performance goals.

Appendix 2. Video Assessment Sheet

Initials (rater): Date:
Subpial Scenario: Practice / Realistic

Video Number:

What did the participant do well?	Identify up to three areas of improvement for this participant:	List two instructions/feedback you would give to this participant:
	1. 2. 3.	1. 2.

OSATS Visual Rating – 7-point Likert Scale

Instrument Handling: how would you rate this participant's ability to handle instruments appropriately and make fluid movements?

Novice 1 2 3 4 5 6 7 Expert

Respect for Tissue: what is the level of care this participant shows for the tissue and the surrounding brain?

Novice 1 2 3 4 5 6 7 Expert

Hemostasis: How would you rate this participant's ability to control bleeding? If no bleeding occurred write N/A.

Novice 1 2 3 4 5 6 7 Expert

Economy of Movement: How would you rate this participant's efficiency of movement?

Novice 1 2 3 4 5 6 7 Expert

Flow: How would you rate this participant's flow of movement in the operation?

Novice 1 2 3 4 5 6 7 Expert

Overall: How would you rate this participant's overall performance in removing a considerable amount of the tumor competently?

Novice 1 2 3 4 5 6 7 Expert