Quantitative Objective Assessment of Preoperative Warm-up for Robotic Surgery

Lee Woodruff White

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:
Blake Hannaford, Chair
Thomas Sean Lendvay
Jay Tal Rubinstein

Program Authorized to Offer Degree:
BioEngineering

University of Washington

Abstract

Quantitative Objective Assessment of Preoperative Warm-up for Robotic Surgery

Lee Woodruff White

Chair of the Supervisory Committee:
Professor Blake Hannaford
Electrical Engineering

Here I present the application of three established methods for quantitatively and objectively assessing robotic surgical performance as well the development and application of a fourth. These four tools are used to assess the hypothesis that a certain surgical warm-up protocol improves performance of surgeons on a da Vinci robotic surgical system. In the protocol, surgeons perform a brief warm-up task on the Mimic dV-Trainer virtual reality simulator prior to performing one of two robotic surgery practice tasks.

Of the four techniques used for performance assessment, the three established techniques consist of basic measures (task time, tool path length, economy of motion and errors), algorithmic assessment (using trained Hidden Markov Model machine learning algorithms) and surgeon assessment (using the Global Evaluative Assessment of Robotics Surgery). The newly proposed technique called Crowd-Sourced Assessment of Technical Skill (C-SATS) draws on crowds of people on the Internet to assess the surgical performance.

The evidence that warm-up improves surgical performance is presented as well as an analysis of the strong agreement between C-SATS and grades provided by a group of surgeons trained to assess surgical performance.

Contents

Page

List of Figures

List of Tables

Acknowledgements

This thesis is complete! I now have time to reflect on the people that made this work possible. Blake, you brought me into the BioRobotics Lab when I felt I had no home at the University, and for that I am grateful. You taught me to invent, patent, publish and transmit our work to the wider world. Tom, working with you was a highlight of my time in the BRL. You offered me the privilege of joining you in the operating room. Your enthusiasm for research along with your dedication to caring for patients inspired me to make surgery the focus of both my research and my career in medicine to come. Both of you offered me mentorship, guidance and an environment in which to thrive. Thank you.

Thank you, to Joan Sanders, Mika Sinanan, and Kristi Morgansen for serving on my Supervisory Committee and to Blake Hannaford, Tom Lendvay, and Jay Rubinstein for serving on my Supervisory Committee as well as my Reading Committee. Your guidance has been invaluable and you've helped tremendously to shape the direction and final form of this thesis.

Tim Kowalewski, I would not have stayed in graduate school if it were not for you.

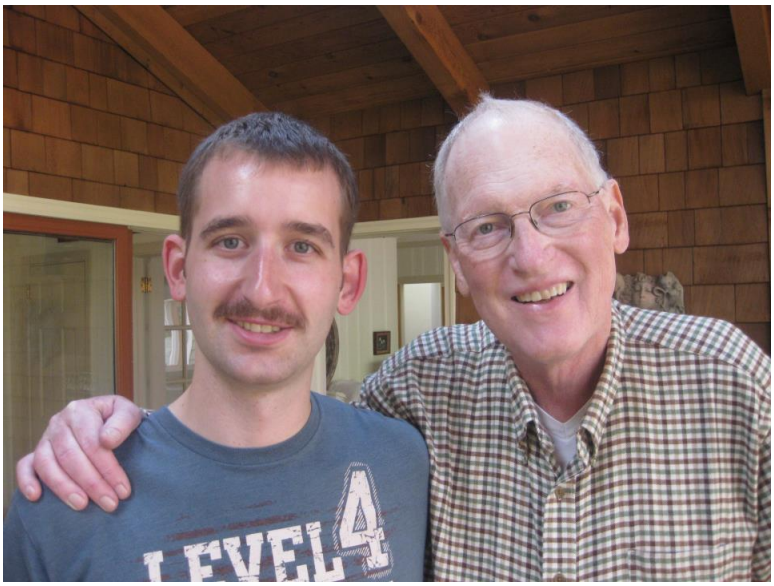
BioRobotics Laboratory family, Hawkeye, Iris, Jack, Andy, Sina, Fredrick, et al. you've been great friends and co-conspirators along this journey. Thank you as well to my friends outside of lab. Floyd, thank you for your advice and for bringing me into the hot shop to work glass. Shaun, you set the bar high. I can't wait to share an OR with you.

To my parents, Jane Grant and Paul White, in additional to the proof-reading of every piece of writing I've ever done, you've also counseled me, supported me, and enabled me to become an

ambitious but compassionate person. To my sister Mariah, you are a great friend and a role model for how to live courageously.

Finally, Shivani, you have been the best companion I could ever hope to have. I am so grateful to you for supporting me through years of graduate school and living on opposite coasts. I am immensely proud of you for your achievements.

Dedication



James Benjamin Grant

August 29, 1945 ~ March 4, 2013

Chapter 1: Introduction and State of the Art

Rapid development and adoption of new surgical devices and techniques may be outpacing the surgical profession's ability to train providers. One recent major new technology is teleoperated robots for surgery (particularly the da Vinci robotic surgical system[1], Figure 1). The safety of patients depends on enabling reliable, safe use of new surgical technology. Virtual reality (VR) surgery simulation is being investigated as a way to train providers for surgery, maximize performance and minimize medical errors. Effective application of new training technology requires sensitive and robust tools to measure surgical performance. A surgeon's skill level and the quality with which they operate vary over many times scales. Over a career, it is expected their average skill increases, but from case to case and day to day their performance may exhibit highs and lows. These variations may be due to environmental factors, patient variation, rest, nutrition, intoxicants, level of training on a surgical system like a surgical robot, time since last use of a system or performance of surgery, etc [1].

Surgery is physically and cognitively demanding, but unlike performers in other demanding areas like sports and performance art, surgeons do not typically warm up for surgery. Recently, researchers have been interested in the use of warm-up tasks including VR simulators to prepare surgeons for the operating room (OR), the hope being that a surgeon's potential performance could be maximized just before the operation begins. There is evidence from other fields to support the hypothesis that warm-up might improve surgical performance but only a few studies have been published to date that quantify its effect specifically on surgery.

[1] Intuitive Surgical, Inc., Sunnyvale, CA, USA

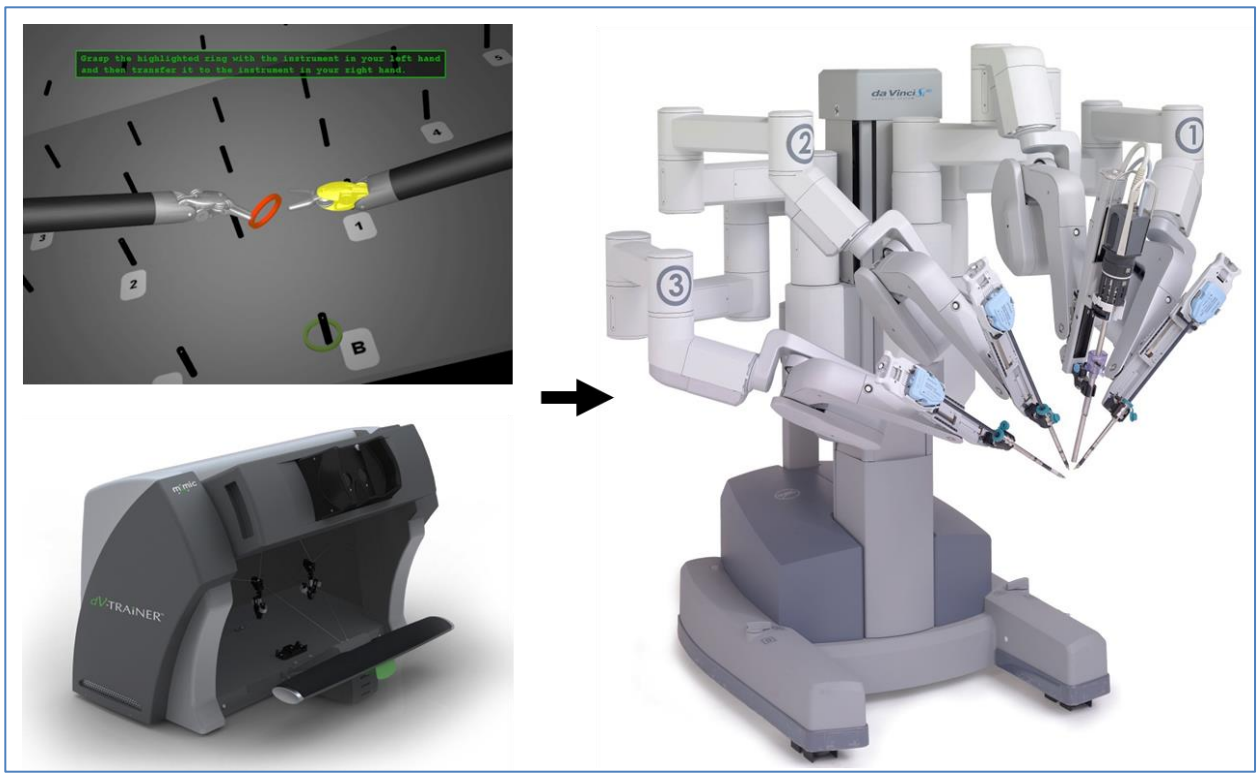


Figure 1 - Will warming up on a dV-Trainer (left) improve task execution on the da Vinci surgical robot (right)?

VR systems have emerged as a valuable training tool for surgery. Skills learned on VR trainers have been shown to transfer to the OR. Currently, VR simulators are available for laparoscopic surgery and robotic surgery, since these domains are inherently performed while viewing the surgical field on a screen, as opposed to with one's own eyes. Robotic surgery VR simulators such as the Mimic dV-Trainer[2] (Figure 1) are well suited for use in the OR as preoperative warm-up. Yet, to date no studies have demonstrated a benefit of warm-up on the performance of robotic assisted minimally invasive surgery, and none have measured the utility of a VR simulator for robotic surgery warm-up. Measuring the impact of warm-up on surgical performance requires valid assessment tools. There are a variety of surgical performance evaluation tools including basic measures (path length, time, economy of motion, mistakes and

[2] Mimic Technologies, Inc., Seattle, WA, USA

errors), structured human assessments (OSATS: Objective Structured Assessment of Technical Skill, GEARS: Global Evaluative Assessment of Robotic Surgery, etc.) and algorithmic assessment using machine learning algorithms such as hidden Markov models (HMM). All of these tools have been shown to correlate with level of training and surgeon seniority and have been adopted as measures of performance quality.

Recently our group developed a new tool to assess surgical performance which is based on a structured assessment tool. This new method, Crowd-Sourced Assessment of Technical Skills (C-SATS) recruits crowds of individuals on the Internet to assess surgical performance and may be a very useful and convenient way to assess surgeons.

Recently our team devised a study to measure the effect of VR warm-up using the dV-Trainer on the performance of robotic surgical tasks on the da Vinci. In this thesis I present the analysis of the data collected in that study. The large dataset we collected was subjected to analysis by basic measures, GEARS and HMMs, and finally C-SATS to test the hypothesis that VR warm-up on the dV-Trainer surgery simulator improves performance of dry lab surgical tasks performed on the da Vinci surgical robot.

In this thesis I present the results of testing the hypothesis that preoperative warm-up improves surgical performance. Confirming this hypothesis through this and other studies may make VR warm-up before practicing robotic surgery standard practice for surgeons. The goals of this thesis are therefore to:

1. Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by basic measures such as task time, path length, economy of motion, and errors.
2. Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by Global Evaluative Assessment of Robotic Surgery.

3. Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by Hidden Markov Models.
4. Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by Crowd-Sourced Assessment of Technical Skill.

This chapter is devoted to describing the context of this research by reviewing the relevant literature. I will describe the evolution of technology in the OR, the need to improve surgical training and the need for tools to maximize the performance of practicing surgeons. I will describe the types of skill used in surgery, focusing on the aspects of performance we hope to measure and improve. I will describe the tools available to assess performance and the evidence for the impact of warm-up on surgeon performance. Lastly, I describe the warm-up study conducted from 2010 to 2012.

1.1 Surgical Robotics

Teleoperated surgical robots were first proposed by Alexander in a 1978 report titled *Impacts of Telemation on Modern Society* [2]. The first use of a robot in surgery occurred in 1985 when a computed tomography guided Unimation Puma 200 robot was used to guide tumor biopsy needles [3]. From the earliest description of laparoscopic minimally invasive surgery, so called *keyhole surgery* has grown into an accepted technique for many procedures [4, 5]. The da Vinci surgical robot was introduced in 2000 and is now used in over 360,000 minimally invasive procedures per year worldwide with an install base of 2,710 robots as of Q1 2013 [6, 7]. In recent years the overall number of medical devices and tools used in the OR has ballooned, each with their own specific set of indications, instructions and operational knowledge. At the same time, the total number of hours a resident physician is permitted to train has been limited

to 80 hours per week [8]. Tools for efficient training of surgeons are needed as are tools to maximize the potential performance of surgeons as they enter the OR.

1.2 Improving Surgery

Despite massive investments in pharmaceutical treatments of disease, surgery has maintained its prevalence. According to physician and public health advocate Atul Gawanded:

> *The average American can expect to undergo seven operations during his or her lifetime. This profound evolution has brought new societal concerns, including how to ensure the quality and appropriateness of the procedures performed, how to make certain that patients have access to needed surgical care nationally and internationally, and how to manage the immense costs [9].*

Medical errors in surgery drive costs higher and result in thousands of injuries and deaths each year [10]. In the year 2000, the Agency for Healthcare Research and Quality reported more than 32,000 deaths resulting from surgery, placing it among the top 10 causes of death in the US [11, 12, 13]. Though not all of these deaths are due to errors, many are and thus may be prevented by reducing the error rate. Furthermore, there are many *adverse events* during surgery which are not considered *errors*. Intestinal perforations resulting in bowel leaks are a known risk of abdominal surgery and while regrettable are generally not considered a medical error. Gawande reasons:

> *Today, surgeons have in their arsenal more than 2500 different procedures. Thus, the focus of recent advances in the field has been less on adding to the arsenal than on ensuring the successes of the treatments we have.*

1.3 Need for Skill Evaluation in Surgery

In many ways surgical success is easily observed. Did the patient survive and thrive following surgery? Was bleeding kept to an acceptably low level? Was a surgical revision required?

However identifying the cause of a surgical error or adverse event is a problem confounded by so many variables that attribution often becomes impossible. According to Gawande "the [New England Journal of Medicine] is entering its third century of publication, yet we are still unsure how to measure surgical care and its results. Experiments in the delivery of care will probably provide the next major advancement in the field of surgery."

During medical school and residency, physicians in training are required to pass the United State Medical Licensure Exam (USMLE). The USMLE is primarily a cognitive test of the subject's knowledge of medicine and its provision. A clinical skills portion of the exam tests subject's ability to interact with test patients but no procedural skills are examined beyond ability to perform a standard physical exam.

In Washington State, physicians are required to renew their medical licenses every 4 years. Renewal requires reporting 200 hours of continuing medical education (CME) [14]. The state does not mandate the content of the CME nor do they require surgeons be subjected to technical skill evaluation.

The American Board of Medical Specialties is an umbrella organization that includes 24 of the 26 medical specialty boards in the United States, including 9 that oversee the training of surgeons. The member organizations such as the American Board of Surgery (ABS) and the American Board of Urology set the educational standards for residency programs providing specialty training in the US. To date, only the ABS requires passing a technical skills exam, the Fundamentals of Laparoscopic Surgery (FLS), in order to attain board certification [15].

The final qualification to perform surgery in a US hospital is hospital surgical privileges. These can be procedure and system-specific. Each hospital establishes their own rules but typically

surgeons must apply for privileges and then perform a number of procedures under supervision of a surgeon with privileges. This requirement is time consuming and imprecise given the variable nature of surgical performance. And there is a clear economic incentive and a potential conflict of interest: hospital profits rely on having surgeons privileged to perform a variety of lucrative procedures. Furthermore, there are open questions as to how to certify a novel procedure and how to translate procedures to institutions that don't currently practice those procedures. Often all that is required to begin using the da Vinci surgical robot in ORs around the country is a robot and completion of a weekend in-service training course provided by Intuitive Surgical. The result of insufficient training can be devastating [16, 17].

1.4 Training Technical Surgical Skills

During their training surgeons develop an arsenal of skills. These include medical decision-making, doctor-patient relationship management, and technical surgical skills. Each draws on a foundation of knowledge, be it of medical facts, psychomotor knowledge or a combination thereof.

> *"Psychomotor learning is the relationship between cognitive functions and physical movement. Psychomotor learning is demonstrated by physical skills such as movement, coordination, manipulation, dexterity, grace, strength, speed; actions which demonstrate the fine motor skills such as use of precision instruments or tools, or actions which evidence gross motor skills such as the use of the body in dance, musical or athletic performance [18]."*

During residency and into professional practice, surgeons develop their psychomotor skills. Out-of-OR practice is growing in popularity as a means of training physicians. VR and phantom tissue model based surgery simulators are commercially available. Practice on these simulators has been show to produce improvements in the OR and they provide new ways to evaluate

surgeon performance [19]. The tasks used to train and evaluate surgeons must be of sufficient difficulty to actually differentiate skill levels. In their analysis of surgeon performance, Rosen et al. found that some surgical tasks (the first step in a laparoscopic cholecystectomy) were easy enough that both novice and expert surgeons performed equivalently [20]. This is similar to the FLS tasks, criticized by some in the field for being too easy. It is argued that even technically deficient surgeons can practice to FLS proficiency.

1.5 Surgical Performance Evaluation Tools

It is believed that medical decision-making and judgment are sufficiently evaluated using written exams. Currently, development and application of tools to measure technical skills, psychomotor skills and surgical tool manipulation skills are of more interest [21]. These skills are also believed to vary over time and be subject to external influences. The following are the existing techniques in use today for evaluating a surgeon's skill.

1.5.1 Direct Observation

William Halstead promoted the apprenticeship model for training surgeons which evolved into the residency training model used today [22]. Direct trainer-trainee interaction allows the attending physician to observe and provide qualitative formative feedback to the resident. This approach has the advantage that it requires no additional equipment, the feedback provided is specific and directed, the trainer has access to patient information and contextual knowledge, and in general this model is compatible with all venues of surgical performance including the OR. Furthermore, feedback can include advice on decision-making.

This approach is limited in that the produced evaluation is inherently subjective and thus inappropriate for summative assessment and certification. Furthermore, the fallibility of human memory and the fact that the assessor can only reference their own personal experiences means that standards will vary across the nation and world. Nevertheless, progression through a board approved residency program is all that is needed to practice surgery today in the US. Residency directors have few tools to prevent trainees from graduating and practicing independently even if they believe the trainee may put patients at risk.

1.5.2 Basic Measures

Basic measures of operative performance for laparoscopic and robotic surgery include task completion time (or subtask completion time), overall path length, and economy of motion (average velocity) [23, 20, 24]. Additional metrics such as tool accelerations and predefined procedural errors also belong in this category. With regards to time, there are definite benefits to minimizing anesthesia, but the correlation between path length and skill is justified more on correlation with seniority (construct validity) than a specific theory of how path length influences patient health [25, 26].

Basic measures are generally very easy to compute. Procedure time for example is routinely recorded for each surgery. These metrics are also considered to be objective. Laparoscopic and robotic systems lend themselves to these types of metrics, indeed the da Vinci surgical robot is internally aware of the position of the end effectors at all times during surgery. Unfortunately, this data is tightly guarded by Intuitive Surgical and made available only to certain preferred research institutions under restrictive conditions. Systems such as SurgTrak,

described below, can achieve end effector tracking but are not yet available for surgery on human patients. There are also general tracking issues, especially in surgical domains such as neurosurgery where the motions are small and influence of tool flexibility is large. Perhaps the most significant limitation of these tools, though, is the fact that there is not an inherent benefit to the patient for their surgeon to achieve the surgery with lower acceleration magnitudes, path lengths, or increased tool velocities.

1.5.3 Objective Global Assessments

A group of Canadian surgeons seeking to measure the surgical skill of their residents first developed the Objective Structured Assessment of Technical Skills (OSATS) in 1997 [27]. Martin et al. at Toronto General Hospital created the tool which uses 7 areas of assessment each graded on a scale from 1 to 5 anchored by text guidelines to assist graders and ensure inter-rater reliability. The seven areas were chosen to represent dimensions of surgical performance deemed relevant to surgical education and patient outcome by the senior staff surgeons. Numerous studies have employed OSATS directly and modified versions of global rating scales to assess surgical performance. Recently Goh et al. created and validated the Global Evaluative Assessment of Robotic Surgery (GEARS) shown in Figure 2 [28]. Their study established the construct validity of the tool, correlating GEARS score during the seminal vesicle dissection portion of a robotic radical prostatectomy to surgeon seniority and training.

Van Hove et al. reviewed recent literature covering OSATS, global operative assessment of laparoscopic skills (GOALS), machine learning approaches and check lists for use in assessing technical surgical skills [29]. They reported each had evidence showing construct validity, the notion that the tool measures what it was built to measure, in this case skill, but that observer

blinding practices were often lax or poorly described [30]. Van Hove makes notes that even a tool without the validation strength to be used for credentialing purposes may be useful as a formative feedback tool for education.

Depth perception				
1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual dexterity				
1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure
Efficiency				
1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression
Force sensitivity				
1	2	3	4	5
Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Autonomy				
1	2	3	4	5
Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate guidance		Able to complete task independently without prompting
Robotic control				
1	2	3	4	5
Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant

Figure 2 - Global Evaluative Assessment of Robotic Surgery by Goh et al. has been demonstrated to be a valid tool to assess robotic surgery.

Global rating scales are popular because of their relative accessibility and ease of use. OSATS and GEARS scores have been shown to correlate with surgeon seniority and cases performed and are often used to assess videos of surgical performances, increasing objectivity of the review. These tools are popular for validating training tools and training curricula.

They are however time consuming to use. Under the best circumstances it takes about the same amount of time to watch a surgical task as it does to assign a global rating scale score. Furthermore, only senior surgeons are trusted to assign global rating scale scores (though crowd sourcing this task is an option). Previous research has shown that clipping, or speeding up and slowing down video of surgical performance influences assigned grades so these practices are to be prohibited [31, 32]. Also, when applied to video recordings of performance, domains such as Autonomy cannot be evaluated and are usually discarded. To date, the presence of a senior surgeon is required to assign scores so they are not available as immediate formative feedback for trainees. When these tools are applied to surgical education research settings, scores from multiple graders are averaged. Scores from single surgeon scorers may not be valid. Furthermore, in current use, the 5 to 7 sub-scores in a global rating scale are summed. This may indicate valuable data is being discarded. Finally, I have been unable to identify any literature that correlates global rating scales with clinical outcomes for patients.

1.5.4 Algorithmic Assessment Using Hidden Markov Models

Markov models and more recently hidden Markov models have proven useful for measuring surgical proficiency [12, 24, 33, 34, 35, 36, 37]. Researchers from the BioRobotics Lab and elsewhere have refined their application to modeling surgical skill over the past 15 years and have demonstrated numerous formulations that are able to correctly group performances into expert and novice categories and assign continuous numerical scores.

The sequence of steps in applying hidden Markov models to performance evaluation typically includes:

1. Capture time varying signals during a surgical task
    a. These often include movement path, tool velocity, tool contact forces, etc.
2. Reduce the dimensionality of captured data to a series of *code words*
3. Train a model
    a. Validate the new model
4. Evaluate a new piece of performance data

Many systems are available for capturing time varying signals. The BioRobotics Lab has used laparoscopic tools instrumented with force/torque sensors and mechanical frames to track the motions of surgeons operating on pigs [24, 38, 39]. The da Vinci surgical robot can provide similar movement data that is sadly locked away from most researchers [12, 23]. Our SurgTrak system, described below, provides similar data about the movement of da Vinci tools commanded by a surgeon. Surgery is fundamentally about manipulation of tissue which requires the application of force but since movement and positioning of tools is also critical and much easier to capture, this is often the basis of surgical skill evaluation, especially on the da Vinci which does not record or report contact forces.

Frequently in surgical skill evaluation, these signals are high dimensioned signals sampled at 10 to 100 Hz. Information content analysis leads us to believe the majority of the data about surgical performance are found between 0 and 5Hz, indicating this sampling frequency is appropriate [24].

Once the data has been captured it must be dimensionally reduced to a series of discreet code words in order to be used to train HMMs. This step is known as vector quantization (VQ). Kowalewski et al. have well described efficient methods to initiate this dimension reduction [40]. Their approach begins with normalizing each dimension of the data by subtracting the mean of each of the data dimensions from the data of that dimension, then dividing each dimension in turn by the range of the data. This range can be the full range of each dimension

or a range that leaves out 2 to 5% of the numerically largest data assumed to be outliers. The result is data that is numerically in the same range. This is important so that numerically large dimensions to not dominate in the next step. Next a k-means algorithm is used to divide the data into n clusters. The value of n is usually in the range of 16 to 256 and is chosen incrementally by finding a value for n that produces a distortion of 1% of the overall distortion of the data, or less than a 1% improvement over a codebook of size 1 larger. Distortion is defined to be the average Cartesian distance between all data points and their corresponding cluster centers assigned using the nearest neighbor algorithm.

Hidden Markov Models are mathematical descriptions of time series systems [41]. They have found successful application in speech recognition and signal processing. They consist of an interconnected set of hidden *m* states that cannot be directly observed. Each of the *m* states can produce one of *n* emissions which are observable. They are parameterized as:

$$\lambda=(A,B,\pi)$$

Where A is an *m* by *m* matrix describing the probability of transitioning from one hidden state to another over one time step. B is an *m* by *n* matrix describing the probability of emitting one of the *n* code words given the underlying state. The vector of length *m* containing the probability of the initial hidden states is signified by π. Not all models include an initial state probability matrix π, instead, it is appended as an additional row on B with 0 probability of returning to that beginning state.

The series of observations or code words are signified as:

$$O = O_1O_2O_3 \ldots O_T$$

and underlying state sequence as:

$$Q = q_1 q_2 q_3 \ldots q_T$$

where the length of each, T, corresponds to the discreet number of time samples in the series. There are three fundamental tasks for hidden Markov models [41]:

Problem 1: Given the observation sequence O and the model λ, how do we efficiently compute P(O|λ), i.e. the likelihood the observation sequence was generated by a system fitting the model. This can be thought of as a "score" or quality factor. Regardless of the actual sequence, the numerical value of P(O|λ) tends to be very small and so is often reported as the log of P(O|λ) and is known as the 'log likelihood'.

Problem 2: Given the observation sequence O and the model λ what is the most likely state sequence Q?

Problem 3: How do we adjust the model parameters of λ = (A,B,π) to maximize P(O|λ)?

The evaluation problem is relatively straightforward and can be calculated in a very short amount of time. The third problem, adjusting the model parameters to fit an observation sequence or set of sequences is more computationally intensive. The two common algorithms are the Baum-Welch Algorithm and the Viterbi algorithm [42]. In each case an initial guess for A and B are provided and their parameters adjusted until the training data fits to a certain quality specification. Guesses for A and B are usually randomly seeded matrices and thus this training task lends itself well to parallelization.

The first problem provides a means of evaluating the fit of a given observation sequence O to a model λ. However log(P(O|λ)) tends to decrease as O increases in size. Rosen's group and the JHU group address this problem in different ways. JHU score each trial against novice,

intermediate and expert models, $\lambda_N, \lambda_I, \lambda_E$, and assigns expertise as a discreet level based on the model to which the user best fits [37]:

$$\text{class} = \text{argmax}(\log(P(O_i|\lambda_N)),\log(P(O_i|\lambda_I)),\log(P(O_i|\lambda_E)))$$

Rosen on the other hand provides a numerical score with a continuous output [20]:

$$\text{Expert Similarity Factor} = \log(P(O_i|\lambda_E)) / \log(P(O_i|\lambda_i))$$

λ_i is a model of surgical performance trained on one's own data. Training this model would take some time but through parallelization may be fast enough to enable near-real time formative feedback to surgeons in training.

Hidden Markov models of surgical skill are distinct from Discreet Markov models (DMM) in that the true underlying state is not known. Rosen has described DMMs where force/torque signatures implied specific surgical motions such as pulling, sweeping, idle, etc. [24]. This approach requires segmented and tagged training data which is very time consuming to produce when analyzing large quantities of data. Automated methods for task decomposition have been proposed for surgical skill evaluation. However, they still require some amount of hand labeling to produce a training set for the classifier [35, 43].

1.6 Warm-Up

Pre-performance practice or warm-up is a popular preparatory activity in many activities from sports to performance art [44, 45, 46, 47]. Benefits include task specific performance enhancement, reduced energy expenditure, reduced rates of injury, and reduction of task time [48, 49]. Bishop et al. reviewed warm-up literature and identified a number of physical mechanisms including increased oxygen consumption, improvement in anaerobic energy

provisioning, reduction in muscle and joint stiffness, and increased nerve conduction rate as contributing to improved performance following warm-up [50]. Bishop also identified positive psychological effects following warm-up. From other studies, warm-up is also known to reduce anxiety and improve cognition [51]. Motor learning literature contributes the notion of motor adaptation. Although still a topic of research, it is known that a user's expectation of the inertial properties of a manipulated object influence the motor commands sent to the muscles [52]. Motor planning adapts to load applied to a user's limbs. This provides the hypothesis that warm-up allows the user to adapt to the mass properties of the master telemanipulator of the da Vinci robot. On our specific case, the user may also be *relearning* the workspace constraints and controls location of the master console.

Preoperative warm-up is being investigated as a way of maximizing the potential performance of surgeons. The first research specifically into warm-up preparations for surgery was performed by Do et al. [53]. In their study 12 residents and 12 medical students performed a laparoscopic transfer task with and without warm-up consisting of repetitions of the same task. It was found that after warm-up residents' pill transfer speed increased by 25% and the medical students' pill transfer speed increased by 29%.

The next researcher to publish an investigation of the effect of warm-up on surgical performance was Kanav Kahol and colleagues [54]. Part one of Kahol's study involved the use of a VR laparoscopy simulator for both warm-up and criterion tasks. 14 post-graduate year (PGY) 1, 10 PGY2, 11 PGY3 and 10 attending surgeons performed two repetitions of the same VR ring-on-pegboard task with the first being marked as the warm-up trial. The VR tasks included both psychomotor and cognitive elements, requiring the subjects to place rings on

pegs by memory after prompting. The reported metrics were gesture proficiency, hand movement smoothness, tool movement smoothness, time and cognitive errors. Warm-up was found to improve performance across all metrics. The study was notable in the use of hand movement and tool tracking as well as the use of HMM based gesture proficiency analysis. Sadly, the group has only published one paper describing their gesture proficiency analysis algorithms and it is not descriptive enough to enable other researchers to try to verify their results [55]. In a second experiment, 6 residents performed VR warm-up tasks followed by a diathermy task on a ProMIS simulator. When compared with a control group of residents, those having performed warm-up exhibited significantly better performance by the same metrics as the first study. This first study is limited by the fact that the warm-up and criterion tasks were identical. The second was limited by the small number of participants and the non-self-controlled design. Both are limited by the opacity of the author's analysis methodology. Calatayud et al. performed the first study examining the transfer of VR warm-up into the OR by measuring the impact of warm-up on the performance of a laparoscopic cholecystectomy [56]. Their study included 8 surgeons and a cross over structure (The original study design included 10 subjects but video recording problems eliminated two subjects' worth of data). The initial subject population included 10 right handed surgeons, half with greater than 100 laparoscopic cholecystectomies each and half with fewer than 40. Half of the subjects performed one cholecystectomy procedure with warm-up and the other half without. Then the two groups switched and the non-warm-up group performed the same procedure with warm-up. Warm-up consisted of 3 tasks on a Lapsim simulator at the medium difficulty level and lasted approximately 15 minutes. Patients were screened to try to assure similarity but patient

variation is not fully controllable. Surgical videos were analyzed with OSATS by expert surgeons. The surgeons' performances were found to be significantly better when preceded by preoperative warm-up, with the warm-up group achieving an average OSATS score of 28.5 out of 35 and the non-warm-up group achieving an average score of 19.25. The Calatayud group describes robust results which are however limited to laparoscopic surgery, a task with fundamental difficulty due to the fulcrum effect [57]. Their results are particularly interesting in that the criterion task was an actual surgery, with scores so strongly in favor of warm-up and effective use of a global rating scale for assessment. They do not report the performance of their subjects on the Lapsim simulator or its utility as a performance predictor.

1.7 Crowd-Sourcing

Crowd-sourcing is the practice of asking large groups of people to perform cognitive or judgment tasks in fields in which they are not trained [58]. Interestingly, in some cases such as solving simple or segmented tasks, crowds of untrained individuals can perform as well or better than experts. Enabled by the Internet, crowd-sourcing has been used for diverse tasks from image analysis to translating images of text into computer-readable text files [59, 60]. In the medical and science realms, crowd sourcing has been used to solve protein folding problems and to offer medical diagnoses [61, 62].

Platforms have been created to harness the power of these crowds. One such platform is Amazon Mechanical Turk (See Figure 3). Mechanical Turk provides a web interface for the crowd (known as Workers and consisting of anyone around the globe age 18 or older and having access to a computer with an Internet connection) to work on tasks presented by

researchers, businesses and others (known as Requesters). The Workers are paid a small commission, on the order of $.05 to $3 for tasks ranging in length from 1 to 10 minutes and including such tasks as classifying items, transcribing audio or video, or any other task the requesters offer. Because Mechanical Turk is a marketplace, the quality and rate of work completion by the Workers is closely tied to the amount paid.

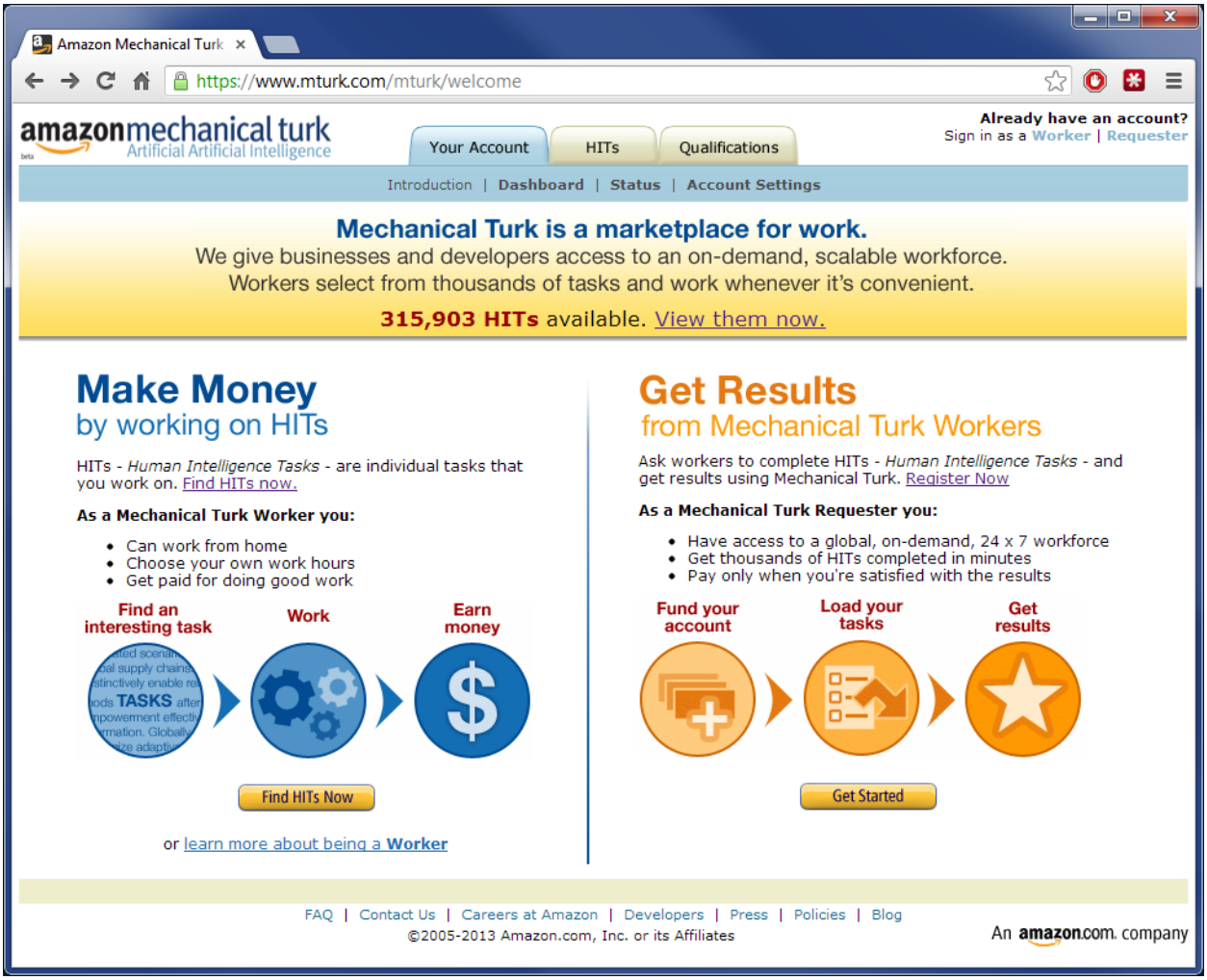


Figure 3 - Amazon Mechanical Turk homepage interface for Workers and Requesters.

1.8 UW/MAMC Warm-Up Study

This thesis involves the analysis of a set of data collected during a study of the effects of VR warm-up on surgical performance. Chapters 2 through 5 all use the dataset generated in this study. This section describes the study and the collection of data. Between September 2010 and January 2012 our group recruited subjects and collected data under Department of Defense Grant W81XWH-09-1-0714: "Virtual Reality Robotic Simulation for Robotic Task Proficiency: A Randomized Prospective Trial of Pre-Operative Warm-up." The objective of the study was to measure improvement in surgical performance on the da Vinci surgical robot derived from a short VR session on a Mimic Technologies dV-Trainer surgery simulator.

1.8.1 Study Tasks

Four physical robotic surgery training tasks were used during the proficiency and primary randomized phases of the study.

1.8.1.1 Rocking pegboard

This was the primary task performed during the randomized portion of the study. The rocker and pegboard are shown in Figure 4. Subjects moved a pair of elastomeric rings with a specified sequence of pegs and tool movements around a pegboard mounted on an S1000-A GyroTwister chemistry rocker[3] undulating at a rate of 8 cycles per minute and with amplitude of ±10° in the roll and pitch axes. It is a novel task based on a VR task used in Kahol's study of warm-up [54]. Mimic Technologies provided a VR version of the task which was used as the

[3] Labnet International, Inc., Edison, NJ, USA

warm-up task for the subjects in the warm-up group. During proficiency testing the task time limit was set to 120% of the average best time of two proficient surgeons participating in the study design. During the primary randomized portion of the trial, the outcome measures were:

- Economy of Motion (continuous)
- Ring Drops (binary)
- Mid-air Transfer Error (binary)
- Out of Order Error (binary)
- Task Time (continuous)
- Peg Touches (counts)
- Cognitive Errors – Mid-air transfer + out of order (counts)
- Path Length (continuous)

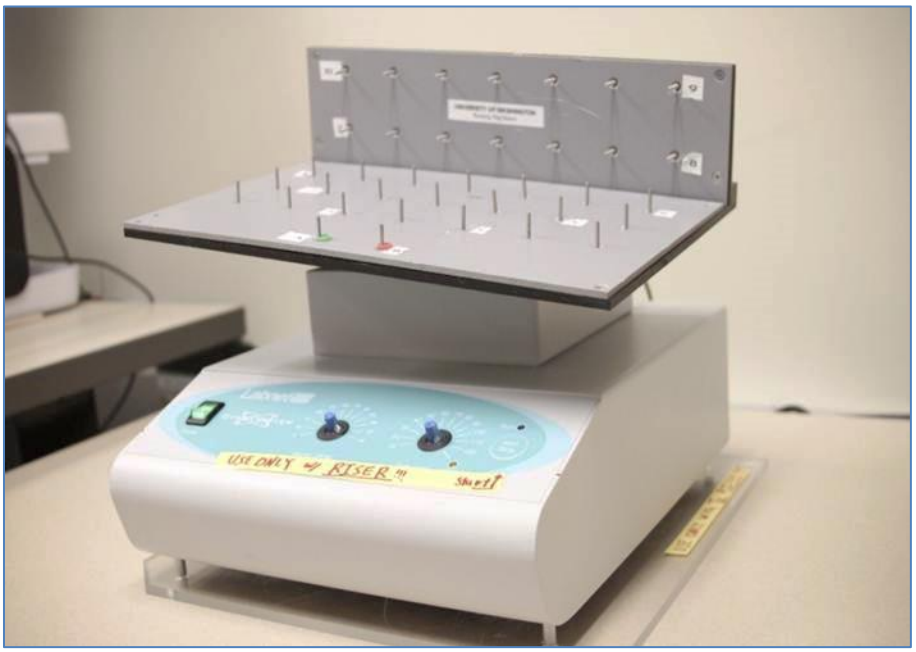


Figure 4 - Rocking pegboard was mounted to a lab mixer rotating at 8 cycles per minute.

1.8.1.2 Suturing With Intracorporeal Knot Tying

This task requires the subjects to drive a needle through a 1.5 inch long piece of penrose drain material and tie a secure surgeon's knot. It is a standard laparoscopy training and evaluation task and is an FLS task. During proficiency testing the task time limit was set again to 120% of a

proficient surgeon's time. During the primary randomized portion of the trial the outcome measures were:

- Entrance Error (binary)
- Exit Error (binary)
- Air Knot Error (binary)
- Break Error (binary)
- Task Time (continuous)
- Economy of Motion (continuous)
- Cognitive Errors – incorrect topology of knot or forgot surgeon's knot (binary)
- Technical Errors – entrance + exit + air knot + break (count)
- Entrance + Exit + Air Knot error (0,1,2,3)
- Path Length (continuous)

1.8.1.3 Peg Transfer

In this task subjects move triangular rubber blocks through a series of motions, first picking up a block from the right set of pegs with the right tool, then transferring the peg in mid-air to the left tool and placing the block on an open peg on the left. Once all six blocks have been moved from right to left they are returned to the pegs on the right, again passing the block between tools. Peg transfer is a standard laparoscopy training and evaluation task and is an FLS task. This task was used only for proficiency testing. The time limit was again set at 120% of the expert surgeons' performances. The task for only used in subject proficiency assessment.

1.8.1.4 Ring Tower

The ring tower task is designed to train the use of the camera clutch and tool clutch on the surgical robot. It involves moving 4 elastomeric rings from a central set of features to a distant set of 4 posts. It is a standard da Vinci robot training task. This task was used only for proficiency testing. The time limit was again set at 120% of proficient.

1.8.2 Study Structure

The study was structured such that each subject had to demonstrate task proficiency on the four tasks: 1) Block transfer, 2) Suturing with intracorporeal knot tying, 3) Ring tower, and 4) Rocking pegboard. Subjects were required to complete two consecutive iterations of each of the proficiency tasks with no errors to be admitted to the study. The subjects were allowed unlimited practice sessions. After a subject had demonstrated proficient use of the robot, they were assigned to either a warm-up group or a non-warm-up group using a four-at-a-time block randomization scheme. Each admitted subject then completed three sessions of rocking pegboard followed by one session of suturing with intracorporeal knot tying, with approximately one to two weeks in between sessions. The warm-up group subjects performed one round of the rocking pegboard task immediately prior to the rocking pegboard and suturing sessions. The non-warm-up group subjects were assigned 10 minutes of pleasure reading. Figure 5 depicts the flow of study subjects through the proficiency and randomized phases of the study.

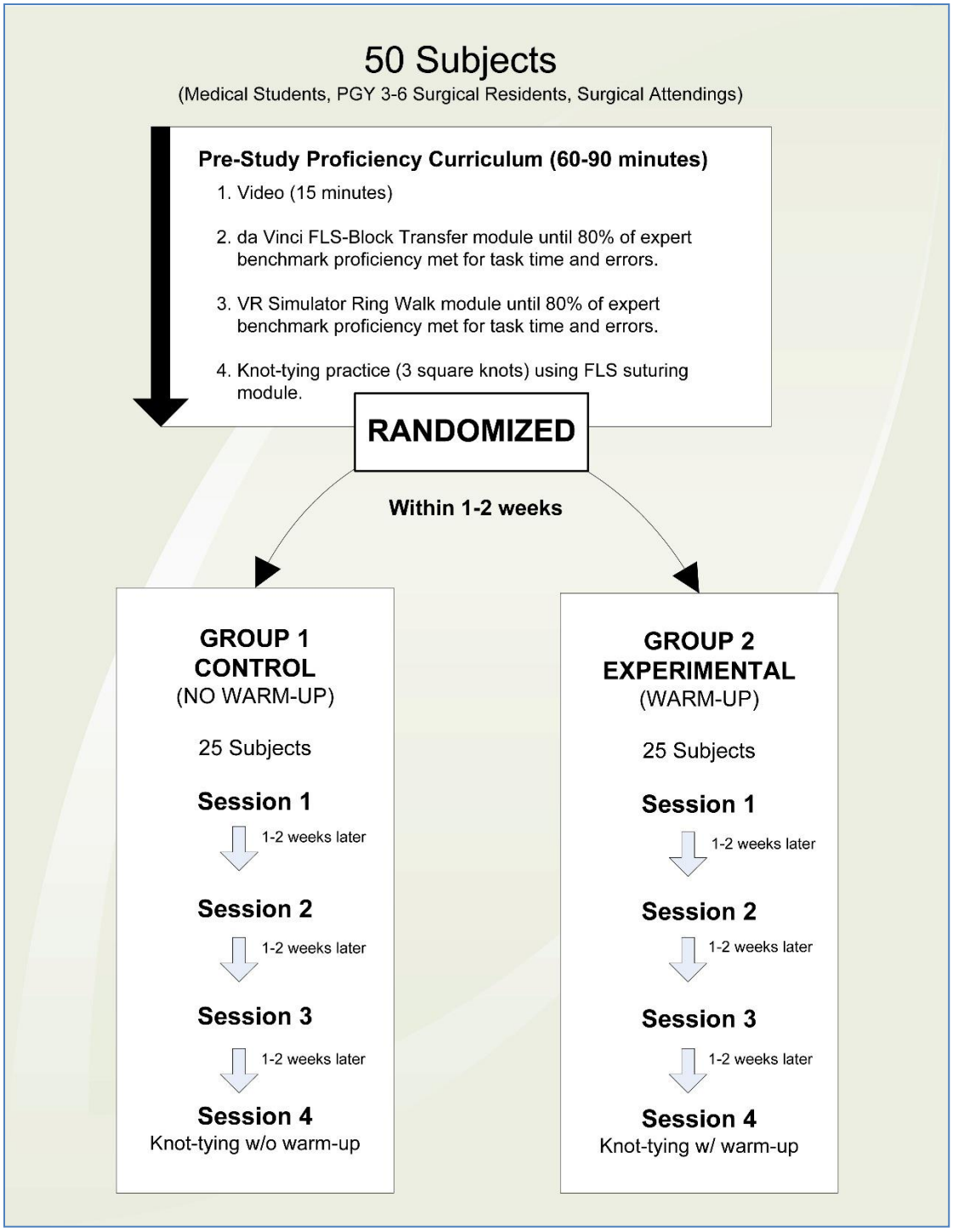


Figure 5 - Flow of subjects through the warm-up study. (Courtesy of Tom Lendvay.)

Table 1 lists the total tasks performed by each subject by the end of the four sessions of the primary randomized portion of the study.

Table 1 - Tasks performed by each subject during the primary randomized portion of the warm-up study. Each subject in the warm-up group performed one round of the VR rocking pegboard task before their robot trials (including before their suturing trial).

Task	Warm-up Group	Control Group
Rocking Pegboard	3	3
Suturing with Intracorporeal Knot Tying	1	1

1.8.3 Demographics

The study was conducted jointly between the University of Washington Medical Center in Seattle, Washington and Madigan Army Medical Center at Joint Base Lewis-McChord outside Tacoma, Washington and included resident and faculty surgeons with and without da Vinci experience. Table 4 in Chapter 2 lists the comparative characteristics of the two groups. They were found to be very well matched and not to exhibit significant differences.

1.8.4 SurgTrak Performance Tracking

We developed a custom system for recording surgical performances on the da Vinci surgical robot. Our system provides surgical performance data locked within the da Vinci combined with endoscope video feed and environmental variables [63, 64, 65]. Figure 6 shows a standard da Vinci Si large needle driver next to a modified SurgTrak Tool. Custom software synchronizes the various data feeds. For the warm-up study, task time, sequence errors, peg touches, position and orientation of the tools, the pose of the tool graspers and surgeon view video were recorded during each task. Table 2 shows the data recorded and its source.

Table 2 - Data types and their sources collected during warm-up study.

Platform	Data Name	Characteristics	Recording Subsystem
da Vinci (recorded with SurgTrak)	Task video	2 dimension left eye view full resolution, contains additional performance data including use of camera clutch and tool clutch	Epiphan[4] DVI2USB
	Tooltip position and orientation	sensor located at back of tool, position and orientation of wrist computed as a known offset from calibration data	Ascension[5] trakSTAR
	Peg touches	Electrical contact between tool tip and pegs (Rocking pegboard task only)	Phidgets[6] Interface Kit
	Grasper pose	Angle of 4 spindles driving the grasper	
dV-Trainer Virtual Reality Surgery Simulator	Task video	2 dimension left eye view full resolution	Epiphan DVI2USB
	position and orientation of tools	end effector location over time	dV-Trainer
	port locations	provided at beginning of task log	dV-Trainer
	peg touches	computed in software	dV-Trainer
	applied force	relative term computed in software but not directly related to a physical force	dV-Trainer

[4] Epiphan Systems, Inc., Ottowa, Ontario, Canada

[5] Ascension Technology Corporation, Milton, VT, USA

[6] Phidgets Inc., Calgary, Alberta, Canada

Figure 6 - SurgTrak modified large needle driver for the da Vinci Si.

Potentiometers were applied to four spindles in the proximal portion of the tool (Figure 7) to measure grasper pose (Figure 8). An electrical contact in each tool detected grasper contact with the grounded metal pegs on the rocking pegboard.

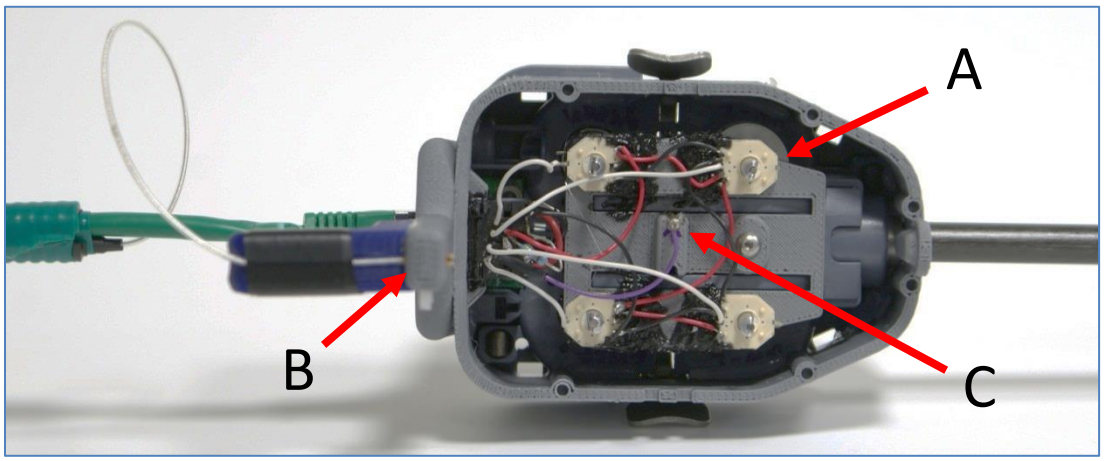


Figure 7 - Internal view of SurgTrak tool including potentiometers to measure spindle angle (A), trakSTAR position and orientation sensor (B) and peg electrical contact sensor (C).

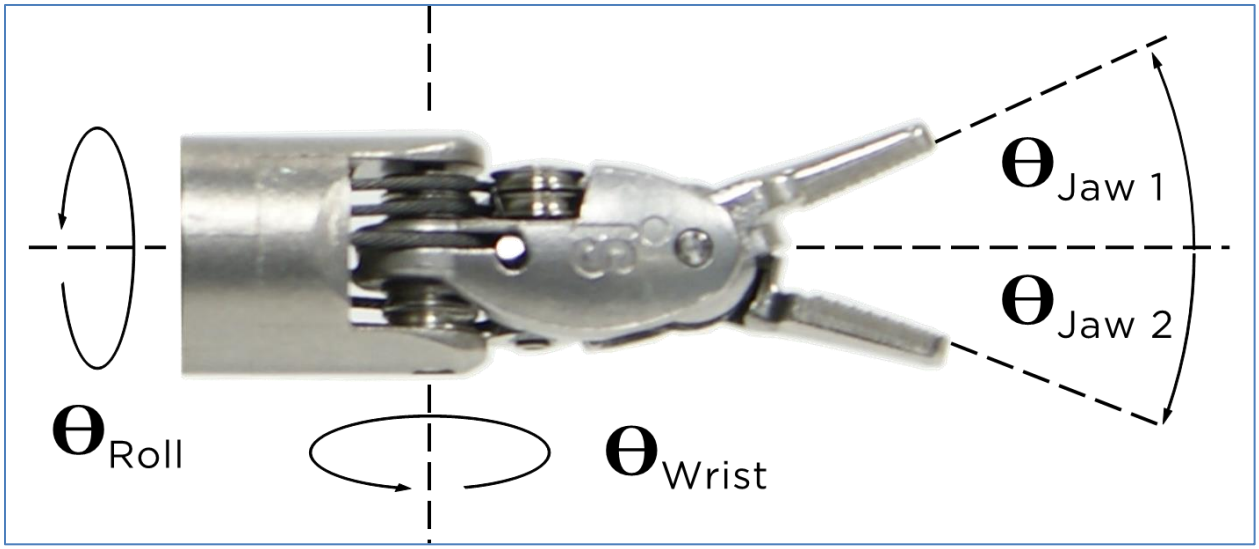


Figure 8 - Depiction of the four spindle cable-driven degrees of freedom of the end effector of a da Vinci large needle driver.

Should force data be determined to be of critical importance, prototype da Vinci tools with integrated force sensors have been devised by other groups [66].

1.8.5 Data Storage

The data collected during the warm-up study is stored on a server housed in the BioRobotics Lab at the University of Washington. The data is password protected. A raw copy of the data is stored with read-only file permissions. A second copy of the data is committed to a subversion repository that tracks all changes to the primary data itself, derived data and associated processing files. This allows for erroneous changes to be reverted. The entire subversion repository was also automatically copied to a geographically remote offsite backup.

1.8.6 Collected Data

The data we collected during the primary randomized trials is summarized in Table 3.

Table 3 - Performance sessions completed during primary study.

	Warm-up		Control	
Subjects	25		26	
Platform	VR	Robot	VR	Robot
Rocking Pegboard	78	78	0	75
Suturing	26 (warm-up task is rocking pegboard)	26	0	25

Chapter 2: Impact of Preoperative Warm-Up on Basic Measures of Surgical Skill

Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by basic performance measures.

2.1 Summary of Contributions

This chapter was published as *Virtual Reality Robotic Surgery Warm-Up Improves Task Performance in a Dry Laboratory Environment: A Prospective Randomized Controlled Study* by Lendvay, et al. in the June 2013 volume of the Journal of the American College of Surgeons [67]. The study drew on the efforts of a large number of researchers. I joined the BioRobotics Lab as the systems to collect the study data were being devised. I designed the surgical tracking technology that became SurgTrak. Along with Timothy Kowalewski, I designed and developed modified da Vinci surgical tools with sensors to track their position and grasper pose, as well as a complementary software program to facilitate recording from the tracking subsystems, and finally a protocol for setting up the da Vinci robot in order to assure optimal data quality. I built many copies of the SurgTrak modified da Vinci tools, making design improvements from my original designs as I went. I helped maintain the data storage infrastructure, trained other personnel to record subject performances, and recorded performances of subjects. We collaborated on writing Matlab[7] software to compute end-effector location using forward kinematics. I debugged this software and built visualizers to verify data integrity. Together we

[7] The MathWorks, Inc., Natick, MA, USA

collated the data from the randomized stage of the study into an excel spreadsheet which we provided to biostatisticians for analysis. I met with our biostatistical experts to help them understand the data we provided them to analyze.

The full citation of the work follows:

Thomas S. Lendvay, Timothy C. Brand, Lee W. White, Timothy Kowalewski, Saikiran Jonnadula, Laina D. Mercer, Derek Khorsand, Justin Andros, Blake Hannaford, Richard M. Satava, Virtual Reality Robotic Surgery Warm-Up Improves Task Performance in a Dry Laboratory Environment: A Prospective Randomized Controlled Study, Journal of the American College of Surgeons, Volume 216, Issue 6, June 2013, Pages 1181-1192, ISSN 1072-7515, http://dx.doi.org/10.1016/j.jamcollsurg.2013.02.012.

2.2 Introduction

Finding methods to improve surgical performance for trainees and practicing surgeons has become a national mission to mitigate surgical morbidity, reduce health care costs, accelerate learning curves, provide curricula for the introduction of new surgical technologies, and ensure that reductions in duty hours for trainees do not compromise surgical education [11, 68, 69]. Surgical simulation methods are mandated by some surgical professional boards [70, 71] and the merits of surgical simulation have been validated both in and out of the operating room (OR) [72, 73, 74, 75, 76].

Most surgical simulation is carried out in dry and animate laboratories at a very different time than actual surgery on patients. But recent studies suggest that surgical simulation immediately before criterion surgical tasks can benefit performance [56, 77]. This presurgical rehearsal, or

warm-up, promises to boost surgical performance. Now that high-fidelity simulator curricula exist for robotic surgery, we hypothesized that virtual reality (VR) robotic surgical warm-up for similar (basic skills) and dissimilar (complex task, intracorporeal suturing) tasks improves performance in both surgical trainees and experienced minimally invasive surgeons.

High-stakes professions, such as athletics and performing arts, have long relied on the principles of the warm-up decrement (ie, the decrease in performance after a period of rest) and the Activity Set hypothesis (ie, the idea that to counter the warm-up decrement, some activity to elevate the arousal and readiness of the subject is required to boost performance) to optimize performance readiness [78, 79, 80, 81]. Yet, surgery does not involve a prescribed warm-up or presurgical rehearsal, although it is a high-stakes profession drawing on intense psychomotor and cognitive efforts. The benefits of warm-up can be particularly important for robotic surgery because of the increased information presented to the surgeon through the visual monitor, as visual cues must be processed to derive forces applied by the tools (synesthesia) and cognitive arousal is likely to benefit greatly from warm-up.

Do and colleagues were the first to use a laparoscopic box trainer to study the effect of warm-up exercises on follow-up laparoscopic tasks, and they observed a considerable improvement in performance (25%) for both residents, irrespective of PGY level, and a medical student control group ($p < 0.0001$) [53]. The study was not able to discriminate the effects of the learning curve vs a true warm-up effect and so Kahol and colleagues sought to address this in a laparoscopic VR simulation study [54]. Surgeons were randomized to receive either warm-up or no warm-up using a series of VR ring-transfer tasks that tested psychomotor, attentional, and visuospatial skills. The results yielded a substantial reduction in errors (33%). In addition, Kahol and

colleagues showed that the warm-up effect was demonstrated in surgeons of all levels of expertise and generalized to dissimilar follow-up tasks, such as an electrocautery task. In Kahol and colleagues' study, both warm-up tasks and criterion tasks were in a virtual laboratory. But in 2010, Calatayud and colleagues showed that a VR simulation warm-up in the OR benefited residents performing laparoscopic cholecystectomies [56]. Eight residents demonstrated higher global performance scoring on an Objective Structured Assessment of Technical Skills tool [56]. In 2012, Lee and colleagues showed that brief reality-based laparoscopic suturing and VR task warm-up immediately before the colon mobilization in laparoscopic nephrectomies performed by senior urology residents yielded higher global assessment scoring and reduced task time [82]. All of these studies have been performed with conventional laparoscopy, yet there have been no studies looking at the value of surgical warm-up in robotic surgery. We sought to explore the role VR robotic warm-up has on similar and dissimilar robotic surgery tasks.

2.3 Materials and Methods

2.3.1 Study Design

Residents and experienced minimally invasive surgery faculty in General Surgery, Urology, and Gynecology from 2 medical centers underwent a validated robotic surgery proficiency curriculum on a VR robotic simulator and on the da Vinci surgical robot (Intuitive Surgical Inc). Once successfully achieving performance benchmarks, each surgeon was randomized to either receive a 3- to 5-minute VR warm-up on the simulator or read a leisure book for 10 minutes before performing similar and dissimilar (intracorporeal suturing) robotic surgery tasks. Three serial trial sessions were performed with similar warm-up and criterion tasks, followed by a

dissimilar warm-up to test generalizability. The primary outcomes analyzed and compared were task time, tool path length, economy of motion, and technical and cognitive errors (Figure 9).

2.3.2 Participant Recruitment

Institutional Review Board approval (#35096) was granted to recruit surgical residents and faculty from the Departments of Urology, General Surgery, and Gynecology at the University of Washington Medical Center and Madigan Army Medical Center to get a representation of both civilian academic and military sector training programs. After acquiring informed consent, each enrollee filled out a demographics questionnaire. The question domains included level of training, handedness, musical instrument and video-gaming experience, and minimally invasive surgery (MIS) experience; all play roles in surgical skill acquisition. All PGY1−6, surgical fellows, and faculty who were experienced in MIS were recruited. All subjects participated in a proficiency curriculum.

2.3.3 Statistical Power/Sample Size Calculation

The statistical power in a repeated measures design was driven by the number of independent subjects in the study, the number of serial observations on each individual, and the degree of within-person dependence among observations contributed by the same individual. Because the within-person dependence was not precisely known, interclass correlations (ICCs) between 0.5 and 0.8 were explored, which covers the typical range for studies involving repeated measurements on the same person [83]. With 51 participants, 3 observations per individual, and assuming an ICC of 0.8, we calculated 95% statistical power for detecting an overall difference between the warm-up and control groups, if the warm-up factor describes at least

an additional 20% of the total variation (0.20 increase in R2). An ICC of 0.8 provides a conservative estimate because it implies observations within subjects will be highly correlated. The statistical power is even higher for smaller ICC values. We did not have preliminary measurements on the path-length metric, however, for the purposes of power assessment, all that matters is the spread of the group means relative to the within-group SD.

2.3.4 Randomization

Permuted blocks randomization was used. Randomization was stratified by site (University of Washington Medical Center and Madigan Army Medical Center) and surgical experience level (resident and faculty). Randomization assignments were provided to each site in sealed, fully opaque envelopes, so that upcoming study group assignments could not be anticipated by study staff or potential enrollees. Randomization occurred at the time the surgeon completed their proficiency curriculum (described later).

2.3.5 Participant flow

Once enrolled, each surgeon went through a robotic proficiency curriculum that included the 90-minute da Vinci online didactics module to familiarize the surgeons with the da Vinci S/Si systems. After passing the tutorial, each surgeon went through a VR (dV-Trainer simulator; MIMIC Technologies, Inc.) and da Vinci dry laboratory robotics curriculum composed of 4 progressively harder surgical skills modules on each platform, respectively (Figure 10 and Figure 11). The proficiency curriculum was generated based on incorporating progressively more complex technical skills, such as object transfer, followed by camera and instrument clutching, followed by all these plus adding motion to the task platform to test spatial relations

capabilities. Proficiency benchmarks were established for each module based on performance by 2 experienced robotic surgeons (TSL, TCB) who have each performed >150 robotic procedures. The benchmark required that each surgeon perform 2 consecutive task iterations within 120% of the mean task time of the 2 experienced surgeons with a zero error rate respective to each module. For example, in the VR Pegboard Level 1 module, a surgeon would have to do as many iterations of the task until 2 consecutive iterations yielded a task time <120% of the mean of the 2 benchmark surgeons performing the same task and with no ring drops or sequence errors. We chose 2 consecutive iterations of success to try to hone the legitimate proficiency of the surgeon for each task. We chose 120% of task time because we did not think it necessary for every surgeon to reach experienced surgeon times to demonstrate proficiency at a particular task. And we did not want to rely solely on task time as the primary benchmark criterion because fast yet error-prone performance is not desirable in surgery.

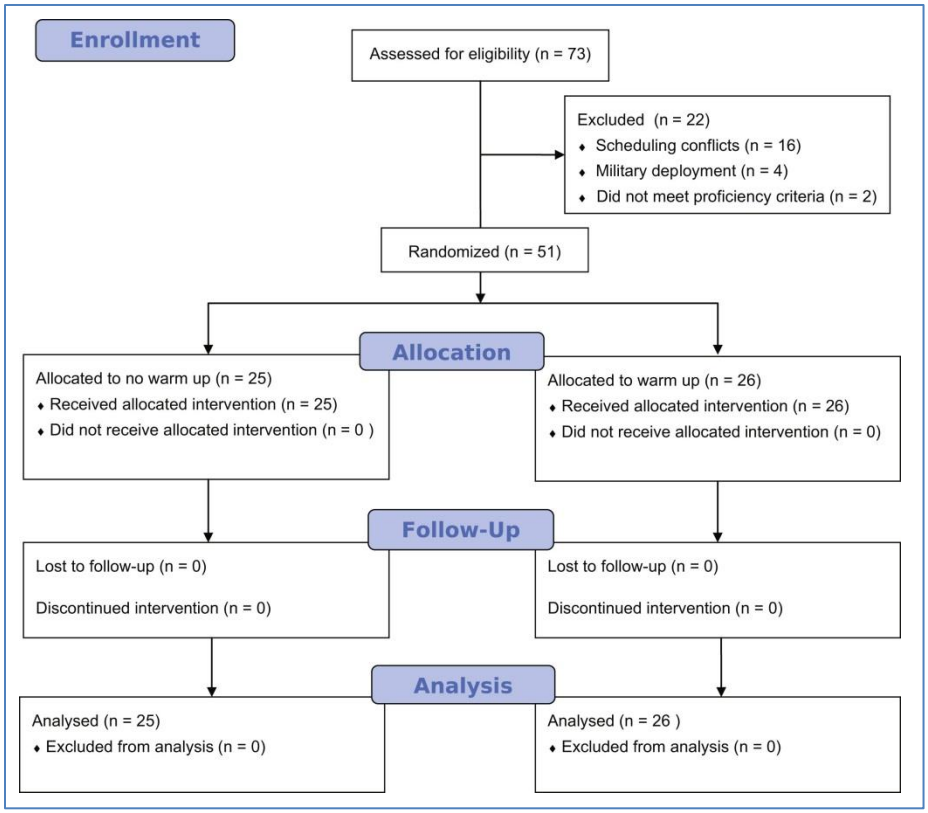


Figure 9 - Patient flow diagram.

Figure 10 - Experimental set-up. Demonstration of proficiency modules in their respective jigs and the plumb lines draping down onto the jig to ensure standard robotic arm positioning.

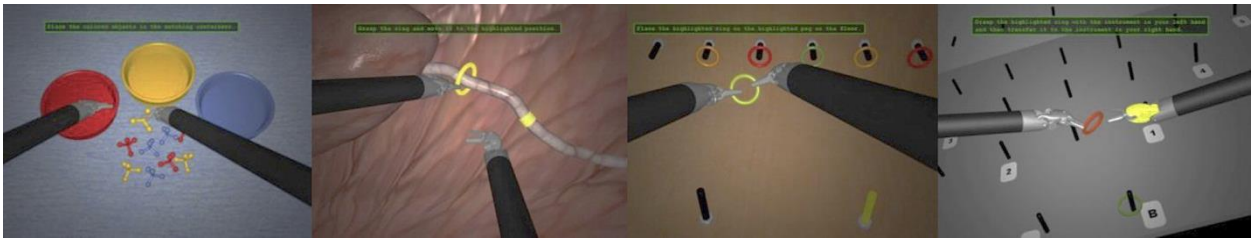


Figure 11 - MIMIC dV-Trainer VR simulation modules from left to right. Pick and Place, Ring Walk Level 1, Pegboard Level 1, Pegboard Level 3.

Concern for learning effect was addressed. To mitigate the confounding effects of the learning curve throughout the study, each surgeon was required to reach proficiency benchmarks before the trial sessions on the da Vinci robot. The intention was to obtain some proficiency equity among the surgeons and familiarity with instrument/camera clutching and manipulation. To equalize the up-front learning of each surgeon irrespective of randomization designation (simulator warm-up or no warm-up), we believed that each surgeon must be given the exact same opportunity to learn the manipulations of both the robot and the simulator to lessen the chance that the warm-up group will have the added benefit of using the simulator at each warm-up trial session.

Once VR robotic simulator proficiency was met, surgeons tested to proficiency on the da Vinci robot through 4 task modules (Figure 12). Construct validation of the da Vinci curriculum was demonstrated through the use of retrofitted da Vinci training instruments capable of tracking tool motions and errors—SurgTrak (described later)—to derive path length and economy of motion performance metrics [84, 65, 85].

Again, proficiency benchmarks had been obtained from the same 2 surgeons and 120% of the mean task times and zero error rates through 2 consecutive iterations were required to advance to the next module. The modules included 2 Fundamentals of Laparoscopic Surgery

(FLS) tasks being performed on the da Vinci robot—block transfer and intracorporeal suturing—because these have been repeatedly validated in laparoscopy curricula [72, 75].

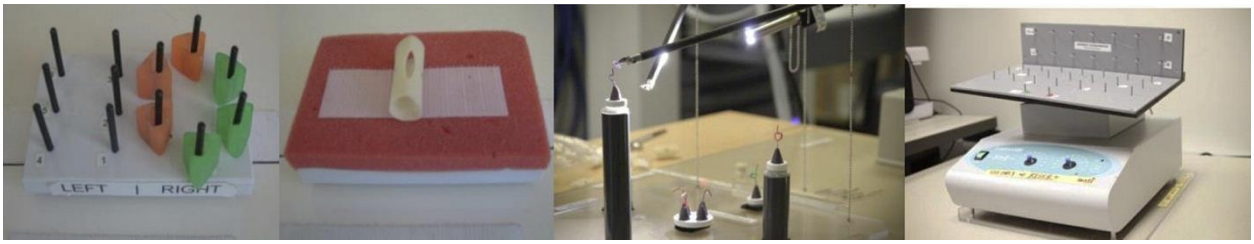


Figure 12 - da Vinci dry laboratory modules from left to right. Fundamentals of Laparoscopic Surgery (FLS) block transfer, FLS intracorporeal suturing (this was the criterion task for session 4), Ring Tower (The Chamberlain Group), Rotating Rocking Pegboard (this was the criterion task for sessions 1 to 3).

2.3.6 Trial Sessions

After a surgeon reached proficiency, he or she was randomized to either the warm-up group or control group. Four trial sessions per surgeon were performed. The first 3 tested performance with or without warm-up on the da Vinci rocking pegboard criterion task. Each of these sessions was separated by a minimum of 24 hours so that one session did not warm-up the surgeon for the next session [86, 87, 88]. In addition, the surgeon could not have performed the trial session if they had done any robotic clinical practice within 24 hours of the session for the same reason. The warm-up group performed the Pegboard Level 3 VR task once directly before performing the analogous da Vinci rotating rocking pegboard task. This generally took 3 to 5 minutes to complete and, unlike in the proficiency curriculum, it was not mandatory for them to perform the VR task with a zero error rate. The controls spent 10 minutes reading a leisure book immediately before performing the da Vinci criterion task so as to minimize the likelihood that they were visually imagining the task to be performed because visual imagery warm-up has been shown to prime surgeons [89, 90]. They could not read any scientific manuscripts or surf the web because we believed that these could also prime the control surgeons.

During the fourth trial session, the warm-up and control precriterion process was the same as the first 3 sessions, but the criterion task became the FLS intracorporeal suturing task to assess whether warm-up generalized to more complex and dissimilar tasks.

2.3.7 Objective Performance Metrics

Based on existing surgical curricula validation studies, we chose the following performance metrics to track on the simulator and the da Vinci robot [76, 91, 92, 93, 21].

1. Total task time (seconds).
2. Cognitive errors (total count): rings placed on incorrect pegs, incorrect sequence of pegs.
3. Technical errors (total count): dropped rings, peg touches.
4. Tool path length (total distance traveled for instruments [mm]).
5. Economy of motion: path length/task time (mm/s).

During the FLS intracorporeal suturing module, additional performance metrics were assessed based on FLS validation of the knot-tying exercise [75].

1. Error: breaking the suture.
2. Error: not placing the suture through the premarked entrance and exit spots.
3. Error: gap left in suture knot (air knot).

2.3.8 SurgTrak Tool Motion Tracking and Video Capture

To capture the objective performance metrics, we developed a system consisting of video recording and surgical tool motion recording combined with custom software. Video was recorded at 30 frames per second from the digital video imaging output of the da Vinci Si/S master console using a DVI2USB device (Epiphan Systems Incorporated). Video was encoded using mpeg-4 compression to produce compact, manageable files. Tool motion data were recorded at 30 Hz. Tool position and orientation were captured with a 3D Guidance trakSTAR electromagnetic tracking system (Ascension Technology Corporation). We retrofitted da Vinci

training tools with rapid prototyped holders for the sensors on the proximal ends of the da Vinci instruments (Figure 13).

Figure 13 - Retrofitted da Vinci training instrument with sensor housing on back end.

These data enabled us to compute path length and economy of motion metrics for each task performance. Grasper pose and electric contact between the tool tips and the pegboard posts were recorded using a PhidgetInterfaceKit 8/8/8 (Phidgets Incorporated). Peg touch errors from the rocking pegboard task were detected and the time of occurrence was recorded by our software. Data streams from the video recording, position recording, and error recording were united using software running on a Windows 7[8]-based laptop computer [65, 85]. Errors on the ring tower, FLS block transfer task, and FLS suturing task were documented in real-time by study personnel and double-checked by video review.

2.3.9 Statistical Methods

Demographic and clinical characteristics measured at baseline were summarized by treatment group and compared with Fisher's exact test for categorical variables and t-test for continuous

[8] Microsoft Corporation, Redmond, WA, USA

variables. The primary comparison for sessions 1 to 3 was a test for the overall mean difference between the warm-up and control groups. Because each surgeon contributed 3 observations to the dataset, this test for continuous outcomes was calculated using a repeated measures ANOVA model and the effect of experience level was investigated by training level (resident vs faculty) and surgical experience (>10 robotic and >10 laparoscopic operations performed as the primary surgeon vs ≤10 cases in each modality). For binary outcomes, repeated measures relative risk regression [94] was used to compare groups and test for interactions. Each surgeon contributed only one observation to the data for session 4 outcomes, t-tests were used to test for a significant difference between study groups, and the effect of experience level was investigated with linear regression models with an interaction term. Session 4 binary outcomes and tests for interactions were modeled with relative risk regression [95]. Data were analyzed using R Version 2.11.1.

2.4 Results

Seventy-three surgeons were assessed for eligibility, with 22 not completing the proficiency curriculum due to scheduling conflicts, military deployment during the study, or inability to meet the proficiency criteria within the study time period. Fifty-one participants, 31 from the University of Washington Medical Center and 20 from Madigan Army Medical Center, were randomized and completed the study (warm-up, n = 26; control, n = 25). Once randomized, no surgeon dropped out. In each demographic category, the surgeons were well matched between the groups, including between faculty and resident participants Table 4.

Table 4 – Demographics and Baseline Characteristics by Intervention Group. (*Comparison of surgeons by group. All categorical variables were compared with Fisher's exact test and age was compared with a t-test.)

Variable	Control (n = 25)	Warm-up (n = 26)	p Value*
Age, y, mean ± SD	35.32 ± 6.47	33.85 ± 5.82	0.40
Sex, n (%)			
Female	10 (40.0)	9 (34.6)	0.66
Male	15 (60.0)	17 (65.4)	
Musical instrument for >3 y, n (%)			
No	7 (28.0)	9 (34.6)	0.76
Yes	18 (72.0)	17 (65.4)	
Handedness, n (%)			
Ambidextrous	0 (0.0)	1 (3.8)	0.36
Left	2 (8.0)	0 (0.0)	
Right	23 (92.0)	25 (96.2)	
Training year, n (%)			
PGY1	1 (4.0)	0 (0.0)	0.34
PGY2	0 (0.0)	2 (7.7)	
PGY3	4 (16.0)	9 (34.6)	
PGY4	3 (12.0)	1 (3.8)	
PGY5	3 (12.0)	1 (3.8)	
PGY6	2 (8.0)	1 (3.8)	
Faculty	12 (48.0)	12 (46.2)	
Subspecialty, n (%)			
Urology	14 (56.0)	14 (53.8)	0.61
General surgery	7 (28.0)	5 (19.2)	
OBGYN	4 (16.0)	7 (26.9)	
Recent video game use, n (%)			
None	15 (60.0)	16 (61.5)	0.99
<2x/week	7 (28.0)	6 (23.1)	
2+x/week	3 (12.0)	4 (15.4)	
Laparoscopic cases, primary surgeon, n (%)			
None	1 (4.0)	0 (0.0)	0.55
<= 10	3 (12.0)	3 (11.5)	
11 - 25	3 (12.0)	1 (3.8)	
25+	18 (72.0)	22 (84.6)	
Robotic cases, primary surgeon, n (%)			
None	9 (36.0)	8 (30.8)	0.61
<= 10	6 (24.0)	10 (38.5)	
11 - 25	3 (12.0)	1 (3.8)	
25+	7 (28.0)	7 (26.9)	

For sessions 1 to 3, testing whether warm-up improved performance with similar VR and criterion tasks, we observed a statistically significant decrease in the task time (−29.29 seconds; p = 0.001; 95% CI, −47.03 to −11.56) and path length (−79.87 mm; p = 0.014; 95% CI, −144.48 to −15.25). Economy of motion favored the warm-up group but was not statistically significant.

Technical errors—dropping rings or touching the pegs with the instruments—did not show statistically significant differences, yet cognitive error reduction favored the warm-up group, but was not statistically significant. The proportion of sessions with errors of placing the rings on incorrect pegs (sequence errors) favored the warm-up group, but because of the wide confidence interval, this was neither statistically significant nor conclusive (p = 0.087;Figure 14; Table 5 and Table 6).

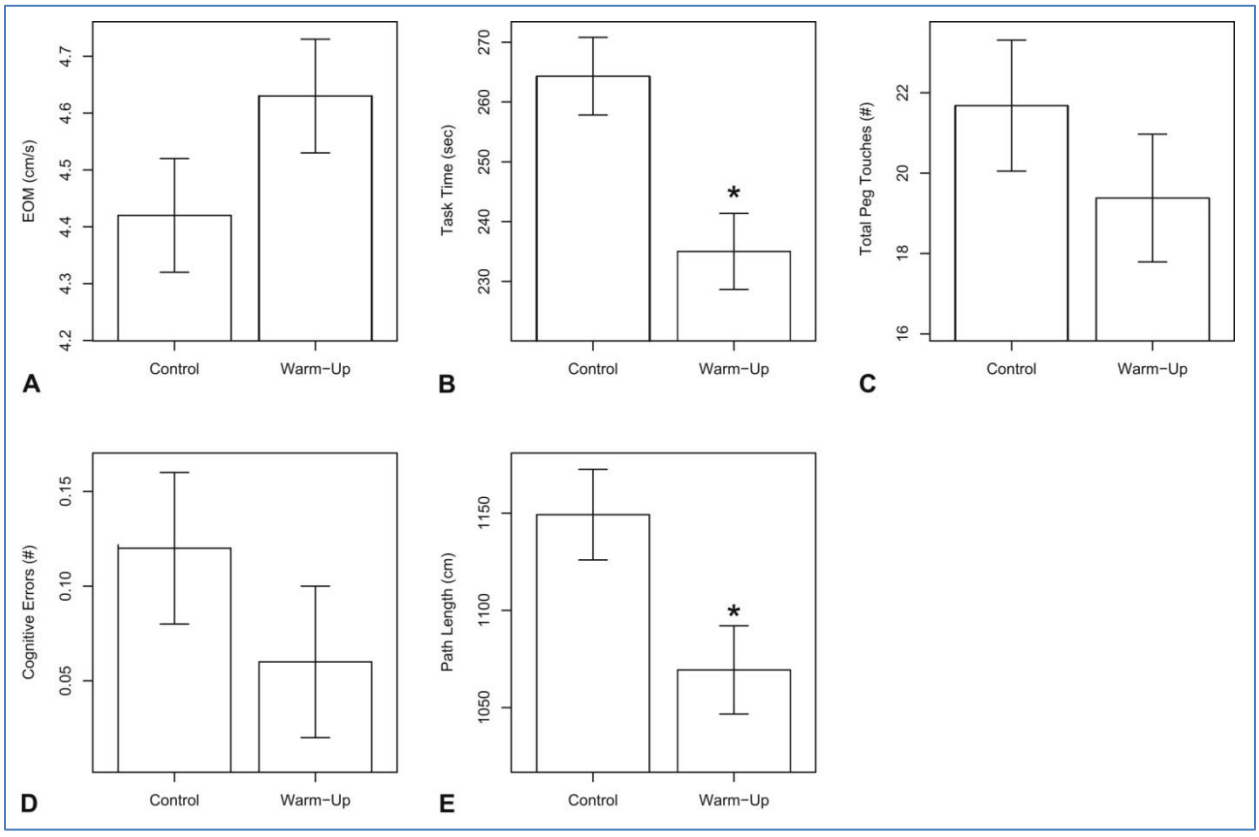


Figure 14 - Control vs warm-up. (A) Economy of motion; (B) task time; (C) peg touch errors; (D) cognitive errors; and (E) tool path length.

Table 5 – Continuous Outcomes by Study Group (Sessions 1 to 3). (Outcomes were individually analyzed with repeated measures ANOVA.)

Outcomes	Control		Warm-up		Difference (95% CI)	p Value
	Mean	SD	Mean	SD		
Economy of motion	4.42	0.66	4.63	0.66	0.21 (-0.06 to 0.47)	0.13
Task time, s	264.31	56.97	235.01	40.11	-29.29 (-47.03 to -11.56)	0.001
Total peg touches	21.68	10.06	19.38	9.01	-2.29 (-6.71 to 2.12)	0.31
Cognitive error	0.12	0.4	0.06	0.3	-0.06 (-0.17 to 0.06)	0.34
Path length, mm	1,149.23	189.03	1,069.37	132.97	-79.87 (-144.48 to -15.25)	0.014

Table 6 – Binary Outcomes by Study Group (Sessions 1 to 3). (All outcomes were individually analyzed with relative risk (RR) regression.)

Error type	Proportion of sessions with error				
	Control	Warm-Up	RR	95% CI	p Value
Ring drops	0.32	0.333	0.96	0.58-1.59	0.87
Air transfer	0.04	0.051	0.78	0.19-3.14	0.73
Out of order (sequence)	0.08	0.013	6.24	0.77-50.76	0.09

For session 4, testing whether a dissimilar VR task can warm-up surgeons for a more complex task (FLS intracorporeal suturing) task, we observed no significant improvements in task time, economy of motion, or path length for the warm-up group. However, when we assessed global technical errors for the suturing (needle entrance, exit errors, and air knot errors, collectively), we observed a near 4-fold reduction in the proportion of sessions with these errors (p = 0.020). Individually, each error was reduced in the warm-up group, but the differences were not statistically significant (Table 7).

Table 7 – Continuous Outcomes by Study Group (Session 4). (All outcomes were individually analyzed with a t-test. *Composite of air knot, needle targeting errors by Fundamentals of Laparoscopic Surgery (Entrance and Exit dots errors).)

Outcomes	Control		Warm-up		Difference (95% CI)	p Value
	Mean	SD	Mean	SD		
Task time, s	111.2	29.3	107.6	37.8	3.6 (-15.4 to 22.6)	0.7
Economy of motion	3.69	0.86	3.82	0.8	-0.14 (-0.61 to 0.33)	0.56
Path length, mm	401.4	114.4	401.5	134.8	0.0 (-71.2 to 71.1)	0.99
Global technical error, count*	0.44	0.58	0.12	0.33	0.32 (0.06 to 0.59)	0.02

When we assessed the effect that MIS experience (>10 laparoscopic and >10 robotic cases as primary surgeon vs ≤10 cases in each modality as primary surgeon) had on the warm-up effect,

we observed that the warm-up effect was more pronounced with experience. Economy of motion (0.63 mm/s; p = 0.007; 95% CI, 0.18−1.09), task time (−53.5 seconds; p = 0.001; 95% CI −83.9 to −23.0), and path length (−97 mm; p = 0.093; 95% CI, −210 to 16) favored the warm-up subgroup among the experienced cohort, and only path length (−75 mm; p = 0.063; 95% CI, −154 to 4) favored the warm-up group in the inexperienced cohort, and not to a statistically significant degree (Table 8).

Table 8 – Effect of Experience on Performance Metrics (Warm-Up vs Control). (The mean (SD) and estimated difference between warm up and control for the 5 continuous outcomes measured in sessions 1 to 3 in the study overall and broken up by robotic/laparoscopic case experience.)

Outcomes	<=10 Robotic and <=10 laparoscopic cases (n = 34)				>10 Robotic and >10 laparoscopic cases (n = 17)			
	Control, mean (SD) (n = 15)	Warm-up, mean (SD) (n = 19)	Difference (95% CI)	p Value	Control, mean (SD) (n = 10)	Warm-up, mean (SD) (n = 7)	Difference (95% CI)	p Value
Economy of motion	4.49 (0.56)	4.51 (0.57)	0.02 (-0.3 to 0.34)	0.9	4.31 (0.78)	4.94 (0.77)	0.63 (0.18 to 1.09)	0.007
Task time, s	258.6 (43.7)	240.8 (39.6)	-17.8 (-39.2 to 3.5)	0.10	272.9 (72.6)	219.4 (41.0)	-53.5 (-83.9 to -23.0)	0.001
Peg touches, counts	24.2 (8.8)	20.7 (9.8)	-3.6 (-8.8 to 1.7)	0.18	17.9 (10.7)	15.9 (6.1)	-2 (-9.4 to 5.5)	0.60
Cognitive errors, counts	0.13 (0.45)	0.05 (0.30)	-0.08 (-0.22 to 0.06)	0.27	0.10 (0.40)	0.10 (0.45)	0 (-0.21 to 0.20)	0.96
Path length, mm/s	1,152 (174)	1,077 (140)	-75 (-154 to 4)	0.06	1,145 (213)	1,049 (118)	-97 (-210 to 16)	0.09

When the groups were divided based on resident (n = 27) vs faculty (n = 24) level, the results were mixed. Path length (−96 mm; p = 0.029; 95% CI, −181 to −10) and task time (−31 seconds, p = 0.013; 95% CI, −55.7 to −6.4) were reduced in the resident warm-up group, and task time reduction only (−27.4 seconds; p = 0.039; 95% CI, −53.5 to −1.4) reached statistical significance in the faculty warm-up group. Path length and economy of motion only favored, but not statistically significantly, the warm-up group (Table 9).

Table 9 – Effect of Training Level on Performance Metrics (Warm-up vs Control). (The mean (SD) and estimated difference between warm up and control for the 5 continuous outcomes measured in sessions 1 to 3 by training level, where faculty are defined as PGY >6 and residents ≤6.)

Outcomes	Residents (n = 27)				Faculty (n = 24)			
	Control, mean (SD) (n = 13)	Warm-up, mean (SD) (n = 14)	Difference (95% CI)	p Value	Control, mean (SD) (n = 12)	Warm-up, mean (SD) (n = 12)	Difference (95% CI)	p Value
Economy of motion, mm/s	4.51 (0.64)	4.69 (0.64)	0.2 (-0.2 to 0.6)	0.35	4.32 (0.68)	4.55 (0.14)	0.2 (-0.2 to 0.6)	0.25
Task time, s	266.6 (51.9)	235.6 (40.5)	-31.0 (-55.7 to -6.4)	0.013	261.8 (62.7)	234.4 (9.4)	-27.4 (-53.5 to -1.4)	0.039
Peg touches, counts	22.1 (10.3)	20.6 (8.3)	-1.5 (-7.6 to 4.7)	0.64	21.3 (9.9)	18.0 (2.3)	-3.3 (-9.8 to 3.2)	0.32
Cognitive errors, counts	0.10 (0.38)	0.12 (0.47)	0.02 (-0.14 to 0.17)	0.84	0.14 (0.47)	0.0 (0.00)	-0.14 (-0.30 to 0.03)	0.10
Path length, mm/s	1,188 (180)	1,092 (140)	-96 (-181 to -10)	0.029	1,109 (192.7)	1,043 (140.8)	-66 (-156 to 23)	0.15

2.5 Discussion

We hypothesized that robotic surgery VR warm-up would enhance technical and cognitive performance on da Vinci dry laboratory tasks. In our randomized study comparing warm-up and control groups of experienced and inexperienced surgeons, we demonstrated that preprocedural warm-up does improve task performance and error reduction. This is a fundamental observation because, to date, the literature has established a warm-up's potential role in conventional laparoscopy, but not in robotic surgery. In addition, laparoscopic warm-up has been shown to decrease operative times in experienced surgeons in the OR, [77] a finding consistent with our observations of warm-up benefiting experienced performers. Many of our tracked performance metrics favored the warm-up group. Task time, path length, economy of motion, error reduction—all surrogates for surgical technical ability—were significantly improved.

We also hypothesized, as Kahol and colleagues showed, that a dissimilar warm-up task can generalize a warm-up benefit or elevate criterion task performance [54]. We observed a statistically significant reduction in the proportion of sessions with global technical errors in suturing, such as air knots and inaccurate needle targeting. The value of this finding is that the ideal warm-up curricula might not need to look like the planned robotic surgery tasks. We did not observe, however, significant improvements in standard technical performance metrics, such as task time or path length, in the generalizability session. It is possible that robotic suturing is so highly technical that psychomotor practice of actual suturing is still the best warm-up task for suturing. When looking at warmed-up urology residents, Lee and colleagues saw a warm-up benefit for a dissimilar intraoperative task of taking down the white line of Toldt in a nephrectomy, but did not see a benefit once the case got to suturing up of the white line at the end of the case. This was explained by the fact that suturing during the nephrectomy was at the end of the case and all surgeons might have experienced the maximal amount of warm-up from all the steps leading up to the end of the case [82].

Similar to the enhancement seen in laparoscopy, we demonstrated a reduction in not only technical errors, but cognitive errors. This suggests that warm-up curricula recruiting not only simple psychomotor centers of the brain, but also spatial relations centers, can be additive to the warm-up benefit. Kahol and colleagues specifically emphasized that warm-up tasks need to not only stimulate psychomotor centers, but also spatial relations and short-term memory centers [54]. In our study, however, some errors were not affected by warm-up, in part because of the low frequency at which these errors occurred, such as peg touches. To observe statistical

significance with this metric, a larger sample size would have been needed, however, it remains unclear whether peg touches are a clinically valid surrogate of precision.

An interesting and unexpected finding was that when the MIS experience of the surgeon was the cohort discriminator, warm-up seemed to benefit the more experienced surgeon. This could be explained by unequal proficiency in robotic skills. We attempted to create a rigorous proficiency curriculum to level baseline robotic skills. And, although all surgeons had met our defined proficiency benchmarks, this most likely did not assure equivalent skills. So we hypothesize that experienced surgeons derive a performance boost from warm-up because they only have to be familiarized with the specific task to do better; they do not have to focus on basic manipulations of the robot itself. Less MIS-experienced surgeons not only require task priming, but may spend additional attentional capacity on performing the basic robotic manipulations (eg, grasping, object transfer, camera and arm clutching). Gallagher and colleagues [76] demonstrated that novice surgeons expend a large proportion of their fixed attentional capacity on performing basic technical skills and experienced performers do not have such high demands on simple psychomotor skills. Experienced surgeons can invest more attention to decision making [76]. These findings about experience are important because there are far more practicing robotic surgeons than there are robotic surgery trainees, and our findings can be relevant to hospital credentialing and maintenance of certification processes.

When we divided the cohort by faculty vs resident, our results were mixed. This might reflect that not all faculty in our group were experienced robotic surgeons because we did not require as an inclusion criterion that “MIS experience” meant robotic surgery experience. Some of

these faculty members had robust conventional laparoscopic experience, but no robotic experience.

There were some key limitations to our study that should be mentioned. First, although we randomized surgeons to 1 of 2 groups—warm-up or control—our proficiency curriculum might not have leveled the proficiency between the groups. Although our groups were very well matched, another design for this study to minimize group skill differences would have been to have each surgeon be their own control.

Second, we strove for intervals between sessions to never occur <24 hours apart or 24 hours from earlier robotic surgery so that one robotic performance did not warm-up the earlier one. However, we do not know if the 24-hour interval extrapolates to robotic surgery. In addition, participants ideally should not have had longer than 2 weeks between sessions, but this was not logistically feasible in some circumstances. Many of our surgeons were on active clinical services and rotated through services that altered the consistency of their intervals. Recognizing the work of Jenison and colleagues, which showed that after 4 weeks of rest, robotic surgery skills degrade, we strove to minimize the number of intervals that exceeded this threshold [96]. We did not, however, adjust surgeons' data based on intervals between sessions.

We have validated portions of the proficiency curriculum using this tracking methodology, but there is potential for varied signal integrity throughout the sessions. The proprietary Ascension software provided us with real-time readouts of the quality of the signal and all our surgeons' sessions fell within the quality requirements of the tracking system, so we believe that we captured accurate data. In addition, task time and error recognition were not dependent on the

tracker data. Signal quality between the transmitter and the sensors on the instruments can be affected by the amount of ferrous material and components generating their own electromagnetic fields. Before enrolling participants, we tested the optimal positioning of the sensors, the transmitter, and the various dry modules to minimize signal distortion. We standardized the positioning of the arms of the robot in relation to the task modules and the transmitter by creating:

1. A jig that housed each module in a fixed position relative to the transmitter (Figure 10);
2. An optimal orientation holder for the sensors on the tools by testing multiple rapid prototyped interface elements before study launch (Figure 11);
3. Plumbs that dangled from set positions on the camera and instrument trocars down to the task module jig that allowed us to set up the robot in identical port configurations between sessions (Figure 10); and
4. Calibration software that tested for sufficient data inputs from all systems before each task iteration commenced.

Alternative instrument tracking methods could include optical fiducials that can be tracked by cameras within the OR, such as those used by Lee and colleagues for their intraoperative laparoscopic study [82]. They tracked surgeon arm and hand movements by affixing sterile markers on the gowns and gloves of the surgeons and used high-resolution cameras to detect precise movements. The advantage of this method is that intraoperative tracking is possible because the markers are sterile, and although the electromagnetic tracker sensors are sterilizeable, the transmitter needs to be within 1 m of the sensors, prohibiting its practicality in the OR. The disadvantage with optical tracking is that this method requires clear line of sight, which is not always possible in the OR. Perhaps a preferable method would be to capture data directly from the da Vinci application programming interface, which has the capability of providing >100 data elements of the instruments' movements in real-time, but such access is

limited to a few centers through contractual agreements with Intuitive Surgical, Inc, and the application programming interface does not capture video or tool contact data [97].

Finally, our findings were unambiguous in a dry laboratory setting, yet the true test of robotic surgery VR warm-up will need to be in the OR, as Calatayud and colleagues and Mucksavage and colleagues did for conventional laparoscopy [56, 77]. This fundamental research in the robotic dry laboratory setting, however, highlights the potential benefit using preoperative VR warm-up for patient robotic surgery to improve patient outcomes and reduce costs. Our experiment used the MIMIC dV-Trainer, which is a desktop platform that has the same VR modules as the current Intuitive backpack simulator that drives VR simulation modules at the da Vinci console. So our findings might be easily translatable into the OR due to the parity between our VR curriculum and what is available today in the OR on the da Vinci Si system. This is a decided advantage for use with robotics systems because the software package that generates the virtual images can reside on any robotic system and, therefore, the preoperative warm-up would actually become part of the operative procedure. Preoperative warm-up in open or laparoscopic surgery, on the other hand, requires an entirely separate simulator to be available in the OR for the surgeon to practice the warm-up. Likewise, in future-generation robotic systems, not only will a warm-up module be included in the robotic system, but downloading patient-specific images (from CT or MRI scans) will also enable the surgeon to perform surgical rehearsal of the critical parts of the operation, so that any errors can be discovered and avoided during the actual operation. The value of mission rehearsal has proven to be of great value in many other domains, such as the military and aviation, and has the potential to greatly increase patient safety in surgery as well [98, 99, 100].

2.6 Conclusions

A brief VR robotic simulation warm-up improves robotic surgery task performance and reduces errors for experienced and inexperienced robotic surgeons in a dry laboratory setting. Further investigation is required to see if these results translate to the OR. These data provide a foundation for future predictive validation studies assessing the role of robotic warm-up for improved patient outcomes and reduced operative cost, and pave the way for novel preprocedural rehearsal investigation in all areas of surgery.

Chapter 3: Structured Surgeon Assessment of Preoperative Warm-Up

Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by GEARS.

3.1 Introduction

The Global Evaluative Assessment of Robotic Surgery has been shown to be a valid tool for assessing surgical performance on the da Vinci surgical robot [28]. It has the potential to be more sensitive than performance metrics based on tool motion alone as the scores can reflect the interaction of the surgeon with the surgical field. We selected a subset of the warm-up data set for scoring. We then recruited 3 surgeons to apply the GEARS system to this set of videos. The impact of warm-up can be detected in these scores. Another use of the GEARS scores is comparison with and evaluation of other methods to assess surgical performance. Ours is the first study ever to measure the impact of VR warm-up on robotic surgery performance.

3.2 Methods

3.2.1 Data

Scoring videos using GEARS can be time consuming and requires surgeons experienced in robotic surgery. These individuals are typically busy people and so a subset of the entire dataset was scored. Table 10 shows the videos scored. 51 subjects proceeded through the study. This should have yielded 51 videos for both the rocking pegboard task and the suturing

task. However, 2 of the 51 performances for the rocking pegboard task and 2 of the 51 performances for the suturing task were not recorded due to errors in the recording equipment. Thus, in both cases, 49 videos were available for analysis. Selecting one rocking pegboard task and one suturing task allowed us to comment on the impact of warm-up when the warm-up VR task was either similar to or dissimilar from the criterion task. Videos from session 3 of the rocking pegboard task were selected as these performances were the most temporarily separated from the proficiency phase of the study. This should minimize the influence of having had recent significant proficiency phase practice on the task.

Table 10 - Tasks scored by expert surgeons using GEARS.

Performance Type	Performances Available for Scoring	Session Number
Rocking Pegboard	49	3
Suturing	49	4

Each of the videos was assigned a unique code that signified the task and subject.

3.2.2 GEARS Assessment Survey Website

In previous experiments we have found that the additional time needed to score a performance using a GEARS tool is negligible relative to the amount of time needed to review the video itself. In order to further minimize the amount of time needed from the surgeons for scoring, we created a web-based scoring system. A home launch page, shown in Figure 15, linked to pages for each scoring 'Task'. Each task page linked to 49 unique scoring pages, one for each performance to be scored. Architecture to automatically generate these pages was created in Matlab. This allowed site revisions to be disseminated across all survey pages quickly. The graphical user interface of the survey was iteratively designed to be acceptable to the graders.

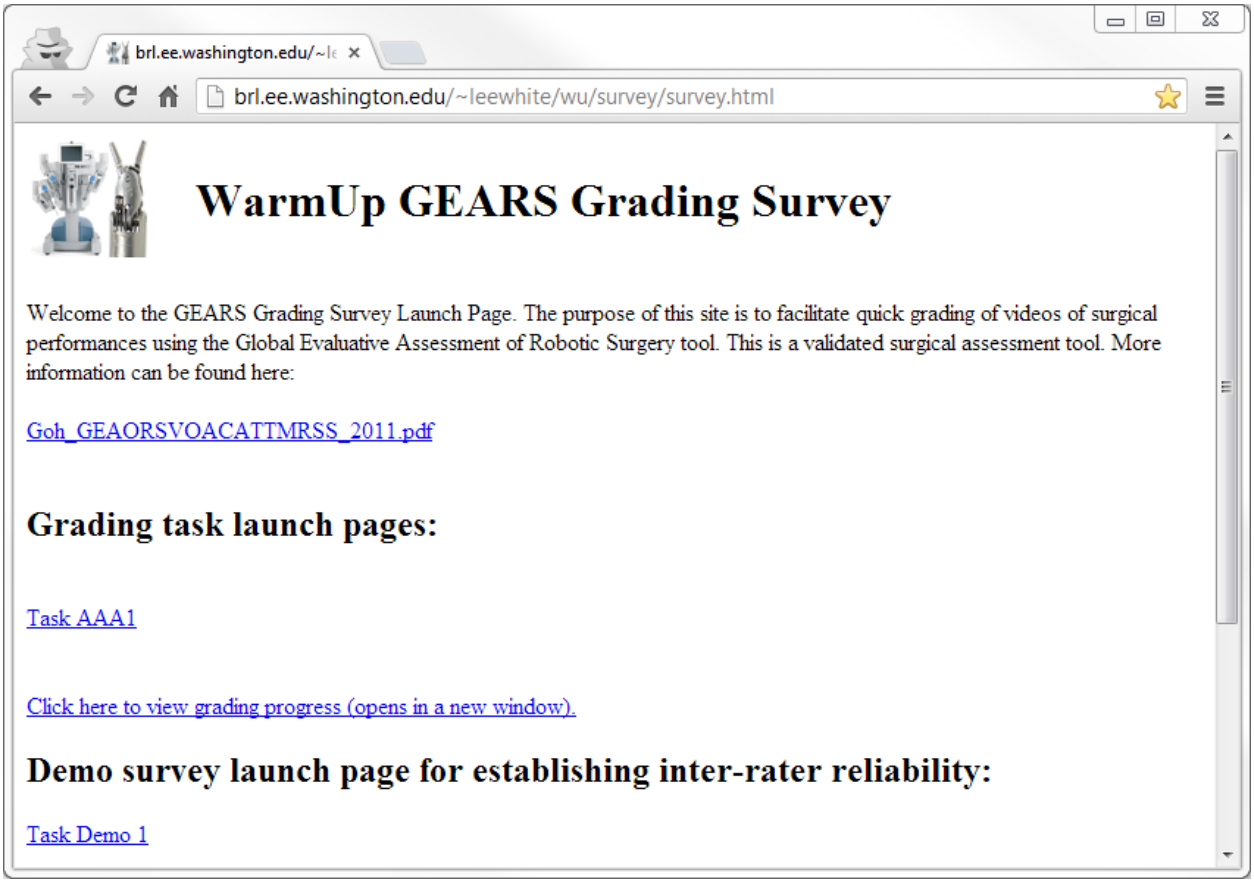


Figure 15 - WarmUp Study GEARS Grading Suite. Automatically generated surveys that are multi-device capable and accessible from anywhere in the world via the Internet.

Figure 16 shows an example survey including a GEARS grading sheet with radio buttons to select the subject's score along with integrated video viewing. The GEARS survey website promotes objective grading by providing a double-blind assessment, where neither the proctor nor the reviewers know the identity or warm-up status of the subject being scored. The fifth GEARS category, Autonomy, was eliminated as the surgeons in this study received no outside direction during the criterion task performances. The survey site was available on the Internet and was compatible with most browsers and devices. This included phones and tablet computers. The surgeon graders were pleased with the convenience of being able to grade anywhere with an Internet connection and at any time during the day, no matter their physical

location. The surgeons were required to enter a username and password when they entered the site. This served a security function and allowed the server to identify the user.

The HTML form website used JavaScript form validation to prevent the surgeons from submitting work before completing scoring of the surgical task. The score data was recorded to a server using the Common Gateway Interface (CGI), a standard Internet method for receiving data from HTML forms. The CGI interface was written in PHP Hypertext Preprocessor. This allowed for some sophisticated quality control features. This flexible and scalable foundation meant the same surveys could be used by small groups of surgeons as well as thousands of graders from across the Internet.

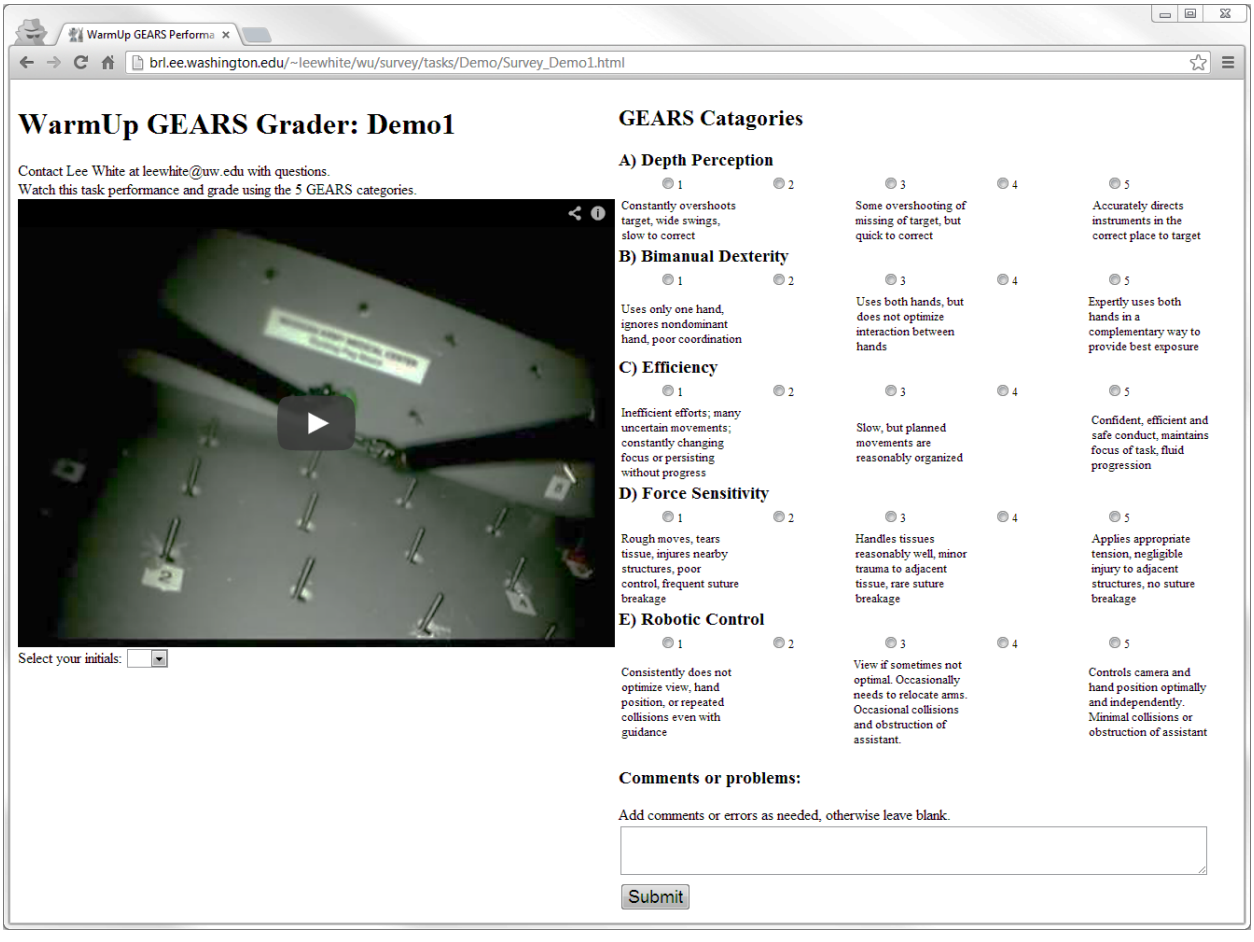


Figure 16 - Example Survey used in the assessment of performance videos, optimized for efficient grading.

3.2.3 Surgeon Scorer Recruitment

A total of three surgeons from UWMC, Seattle Children's and MAMC were recruited as scorers. Table 11 shows the overall level of experience of the recruited surgeons.

Table 11 – Surgeon graders' training and experience.

	Years as Attending Surgeon	Years as Robotic Surgeon	Estimated Cases on da Vinci
High	8	7	500
Low	4	6	150
Mean	6	6.66	270

3.2.4 Grader Selection and Assurance of Inter-Rater Reliability

The textual anchors included in the GEARS scale are meant to assure graders understand the scoring criteria (see Figure 2). This was intended to promote consistency between raters, known as inter-rater reliability [27, 101]. Inter-rater reliability helps ensure the validity of assigned performance scores. Scorer agreement of 0.8 computed using Cronbach's alpha indicates "Good" agreement. We have found that scoring criteria alone are not enough to ensure acceptable levels of agreement. A better way to assure inter-rater reliability is to have the graders practice by reviewing a sample set of data. Prior to grading the main corpus of data, the three graders completed 10 demo surveys for tasks from the proficiency phase of the study. Then a teleconference was held to compare their scores and re-watch the videos from the surveys they completed.

3.2.5 Statistical Analysis

Comparison of warm-up subject performance vs. control subject performance was computed using a non-paired single tailed t-test at the 0.05 significance level [102]. This test is appropriate because the subjects in the two groups are different. The hypothesis that warm-up

produces an improvement in performance, and thus GEARS score, makes a single tailed test appropriate. We rejected the null hypothesis of no warm-up effect if the test statistic computed on the collected normalized scores for all subjects is less than 0.05.

We anticipated that some of the variance in surgical performances would be due to overall skill and level of training. Senior surgeons in the study had a better average performance than their resident counterparts, regardless of warm-up status. For this reason the groups of performances scored using GEARS were also divided by level of training. This also allowed us to see if warm-up affects experts differently than novice surgeons. The 'expert' group criteria was that a subject must have performed more than 10 robotics cases as primary surgeon and more than 10 laparoscopic cases as primary surgeon. All subjects not meeting these criteria were deemed 'novice'. This subdividing of data does put us at risk of failing to have the statistical power to detect an improvement due to warm-up

3.3 Results and Discussion

3.3.1 Expert Surgeon Graders Calibration to Establish Inter-Rater Reliability

The group of three experienced surgeons achieved an initial agreement of 0.7620 when grading 10 videos presenting a range of skill values. This first round of grading also served to expose the graders to the GEARS grading website and to expose them to a range of performances. After the grader calibration meeting, the surgeons felt they had come to an understanding of one another's grading expectations and so we decided to proceed with grading the first round of performances, the rocking pegboard task. Grader agreement improved, as shown in Table 12, with that task and improved further still when the surgeons graded the suturing task. It may

be because the surgeons grading performance improves with experience or perhaps that the suturing task tends to be more accurately graded using GEARS. Another possibility is that the shorter duration of the suturing task, generally less than 2 minutes as opposed to 3 or more for rocking pegboard, led to a better experience for the graders. Interestingly some of the graders saw the grading task as entertaining while others saw the work as arduous. The surgeons began grading the main body of work on April 30th, 2013. The last surgeon finished the final grading task on May 29th, 2013, 29 days later. In that period of time each surgeon graded 98 performance videos. The actual amount of time spent grading was of course much less. The surgeons reported finding time to devote to grading to be the primary challenge.

Table 12 – Surgeon grader agreement within certain tasks.

	Rocking Pegboard	Suturing
Cronbach's alpha agreement	0.7876	0.8879
Confidence	Acceptable to Good	Good to Excellent

Figure 17 shows the overall agreement between individual graders. The graphs include each of the 49 rocking pegboard and 49 suturing scores, with mean scores as well as scores for individuals graders represented. The performances are sorted in order of increasing mean score. It can be seen that while the graders express the same trend, the third grader often scores higher than the average and grader 1 often scores lower than the average score. One interpretation of this finding is that in situations where only one grader is available, a 'handicap' could be applied to scores they provide to try to recover a more accurate performance assessment. Such a handicap could be computed after having the individual grader assess a standard set of 'gold standard' data for which a true score is known.

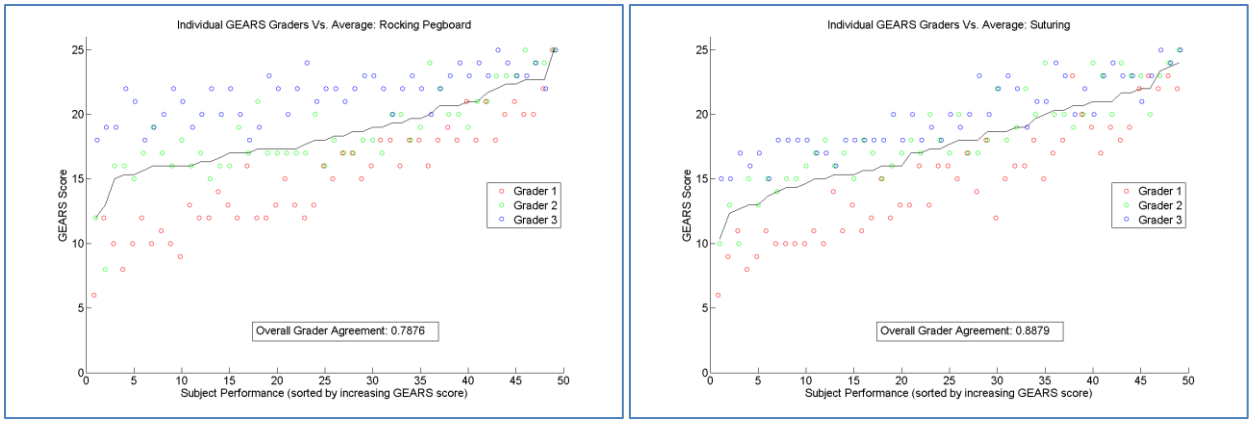


Figure 17 – Agreement between surgeon graders. LEFT: Rocking Pegboard. RIGHT: Suturing. Performances in each task are sorted left to right by increasing average GEARS score.

3.3.2 Influence of Warm-Up on Robotic Surgery GEARS Scores

When considering the impact of warm-up on the two tasks, we first look at the performances as a group. As seen in Figure 17, for the rocking pegboard task the warm-up group achieved a mean GEARS score of 17.63 while the control group performed slightly better, scoring 17.68 on average. For the suturing task the warm-up group outperforms the control group 18.92 to 17.89. However, in neither case do the differences between the distributions of scores achieve statistical significance.

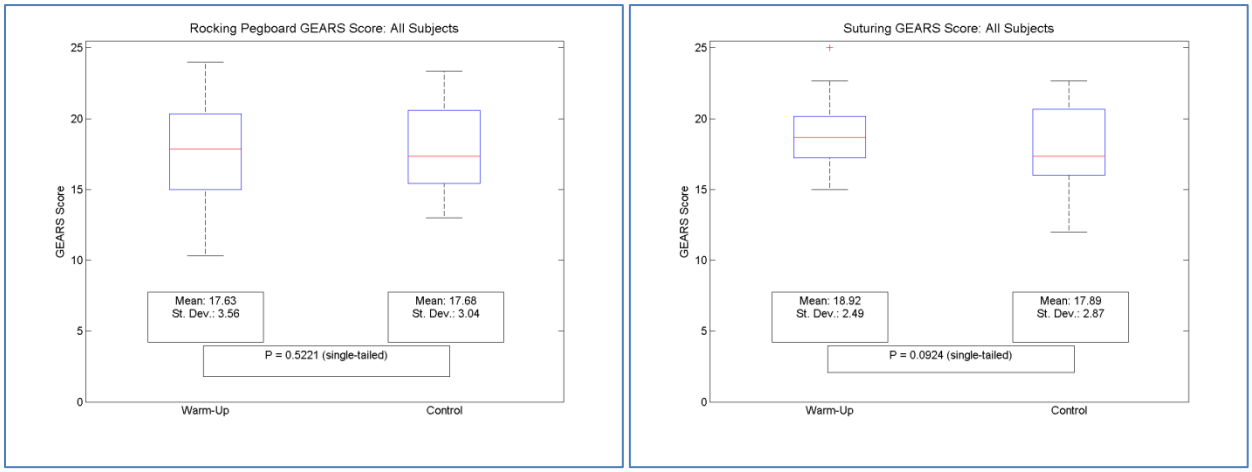


Figure 18 – GEARS Scores for All subjects. LEFT: Rocking Pegboard. RIGHT: Suturing.

When we separate the groups into experts and novices and consider their performances separately, we find that in all but one case (novices on the rocking pegboard task), GEARS scores favor warm-up, however only for experts does this difference approach statistical significance at the level we selected. These results are depicted in Figure 19 and in Figure 20 as well as summarized in Table 13.

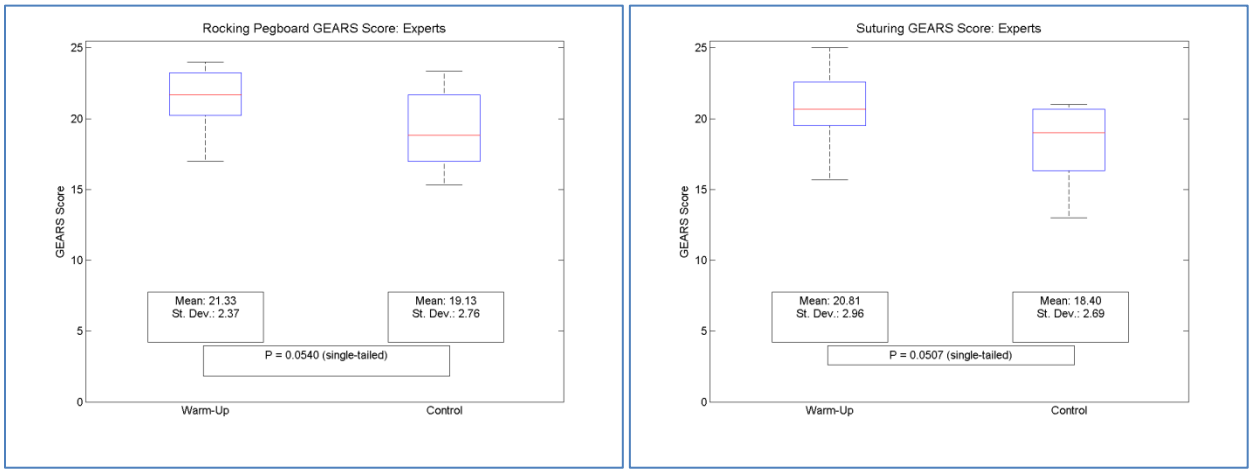


Figure 19 – GEARS Scores for Experts only. LEFT: Rocking Pegboard. RIGHT: Suturing. In both tasks the GEARS assessment nearly demonstrated statistically significant differences.

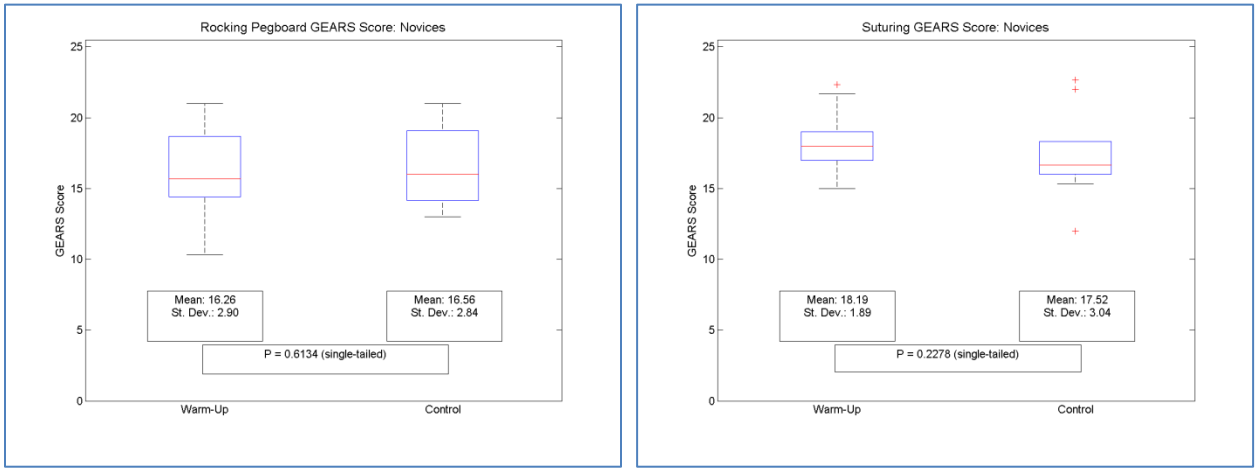


Figure 20 – GEARS Scores for Novices only. LEFT: Rocking Pegboard. RIGHT: Suturing. For rocking pegboard, warm-up seemed not to improve performance. For suturing, warm-up did increase the mean score but this difference did not rise to the level of statistical significance.

Table 13 –Warm-up impact on surgeon GEARS scores.

Group	Task	Warm-up	Control	t-test
All Subjects	Rocking Pegboard	17.63	17.68	0.522
Experts	Rocking Pegboard	21.33	19.13	0.054
Novices	Rocking Pegboard	16.26	16.56	0.613
All Subjects	Suturing	18.92	17.89	0.092
Experts	Suturing	20.81	18.40	0.051
Novices	Suturing	18.19	17.52	0.228

3.4 Conclusions

Taken as a whole, it appears that GEARS scores do favor warm-up over the control group. However, as was the case with basic measures, the improvement is slight. This is likely due in part to the non-cross-over design of the study. We are not able to compare a subject's performance with warm-up to their performance without. While others such as Calatayud in particular observe an improvement of 9 points in a structured assessment tool following warm-up, we were unable to measure such an improvement. This may be due to the nature of the warm-up phenomena, the design of the study (Calatayud utilized a cross-over study design), the applicability of warm-up to robotic surgery, some feature of how we implemented the warm-up task or another factor.

Chapter 4: Hidden Markov Model-based Assessment of Surgical Motions Following Preoperative Warm-Up

Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by Hidden Markov Models.

4.1 Introduction

Using algorithms to assess surgical performance has been a goal and focus of research since methods to track surgical motions were first created. Hidden Markov Models (HMM) are the most popular method for encapsulating the characteristics of surgical performance that track surgical skill. The appeal of skill assessment algorithms is obvious. Once a model is created, assessing a surgical performance with an algorithm costs next to nothing. This makes these methods attractive to those performing studies of surgical training curricula since such studies often involve the assessment of large numbers of surgical performances. These algorithms may one day be used for surgical credentialing purposes. The methods to apply HMMs to surgical performance are fairly well established but even those wishing to use them without developing new types of models must build their models carefully for them to be useful. We decided to employ HMMs in the assessment of the performances of the warm-up study to evaluate the warm-up hypothesis.

4.2 Methods

4.2.1 Data

In order to be able to use expert surgeon scores for comparison with scores produced using HMMs the same set of tasks were used in this phase of the study as in the GEARS assessment phase. These performances are listed in Table 14. Of the original 51 subjects, problems with data capture eliminated 4 and 5 suturing performances and rocking pegboard performances, respectively, from our analysis.

Table 14 - Tasks assessed using HMMs.

Performance Type	Performances Scored	Session Number
Rocking Pegboard	47	3
Suturing	46	4

The raw data fed into the analysis and used to train the models included the Cartesian position of each tool as well as the four spindle angles that determine the grasper pose of each tool. This made for a total of the 14 continuous variables listed in Table 15. SurgTrak software recorded the tool motions at a rate of 30 Hz.

Table 15 – Variables used in the training of skill assessment HMMs.

Source	Units	Details	Dimensions
Cartesian Position	Inches	X, Y, Z position	3 per tool
Spindle Angles (grasper pose)	Radians	Left Jaw, Right Jaw, Wrist Angle, Tool Roll	4 per tool

4.2.2 Data Processing

It is standard practice to convert the data streams to velocity. The justification being that since surgical targets can vary in location through a patient's body, the absolute location should not

be considered. For this reason the velocity of each variable was computed using standard numerical methods.

The resulting velocity stream from each of the individual performances from either the rocking pegboard or suturing task were concatenated and then normalized by computing the mean and standard deviation in each dimension. The associated mean was subtracted from each data stream. The result was divided by the standard deviation associated with that dimension. This resulted in a 14 dimension block of velocity data, the length of which was the total number of samples across all performances of a given task type. The channel means and standard deviations computed before are essentially the channel scale factors and so they were saved for later encoding of data.

4.2.3 Vector Quantization

The two data blocks produced by the above processing were then clustered using a k-means algorithm. The k-means algorithm provides a way to reduce the dimensionality of a data set, in our case, from 14D continuous at 30 Hz to a 1D discreet variable at 30 Hz. The k-means algorithm takes as its input a block of data to be grouped into clusters and the desired number of clusters. From the data set a group of k points are selected at random. These become the cluster centers. All point closest to a given cluster center are assigned to that center. The aim of the algorithm is to iteratively move these cluster centers so as to minimize the sum squared distance from each point to its cluster center. The optimal number of clusters, k, is determined by running the k-means algorithm with varying cluster sizes and choosing the cluster size that generates a 1% improvement in distortion of the data over having k-1 cluster centers, as measured by the average sum squared distance from the each point in the data set to the

nearest cluster center. Table 16 shows the optimal number of cluster centers for the two tasks in this study.

Table 16 – Computed optimal codebook sizes.

Task	Size
Rocking Pegboard	70
Suturing	55

Finally, data were clustered with the optimal number of centers shown in Table 16. These cluster centers, collectively known as the *codebook*, signify the k *codewords* that summarize the dataset. Each performance in Table 15 was scaled using the task-specific scale factors and encoded using the appropriate generated codebook. The nearest-neighbor search algorithm is used for this task and produces 1D by 30 Hz 'observation sequences' consisting of the index number (1 to k) of the cluster center closest to a given point in the performance. These observation sequences can be used to train HMMs.

4.2.4 Hidden Markov Model Training

For algorithmic assessment in this study we used a 15 state black-box HMM. This type of model is consistent with the most successful vetted models chosen by Rosen et al. in their latest work [24] and by Kowalewski in his thesis [103]. In both cases laparoscopic instead of robotic motions were analyzed and in both cases force data was also used in model training. In our approach we do not have force data. To train the models, 35 randomly initialized models were generated. These models consisted of a 15 by 15 state transition matrix and a 15 by k emissions matrix. Selected training data was applied to the models and the models were allowed to train until a tolerance of 0.0001 was met. The training iterations were limited to 500 and in the rare case that a training failed to converge, the model produced was discarded. Of

the generated candidate models, the model with best fit to the training data was selected for further use. The fit criteria was:

$$argmin_{j=1}^{35}\left\{\sum_{i=1}^{n}\log(P(\mathrm{O}_i|\lambda_j))\right\}$$

Where λ are the candidate models and the set of n observation sequences was signified by O.

4.2.5 Scoring of Performances and Statistical Analysis

A continuous numerical score was required for each subject performance. All of the performances were scored using the following statistic similarity factor formula presented by Rosen [34]:

$$\text{Expert Similarity Factor} = \log(P(O_i|\lambda_E)) / \log(P(O_i|\lambda_i))$$

Thus to score performances, two models were needed: one for experts and one for novices. We used the same single tailed t-test to compare the expert similarity factor performance scores of the subjects in the warm-up and control groups based on the same justification discussed in the Chapter 3. The null hypothesis will be rejected if the test statistic computed on the data is found to be below 0.05.

4.2.6 Selection of Model Training Data

Recently Tim Kowalewski of the BioRobotics lab came upon a new notion for choosing the surgical performances with which models should be trained [103]. Even very experienced surgeons perform poorly on occasional tasks. Since the HMMs will incorporate any data on which they are trained, it was reasoned that only the best runs of the most experienced surgeons should be used as model training data. Furthermore, it was reasoned that all runs

where the surgeon performs an error should be excluded. Figure 21 shows how the expert and novice training trials and were selected for each task. The performance models were trained with data from the same task as the criterion task, i.e. rocking pegboard performances will be evaluated with rocking pegboard models.

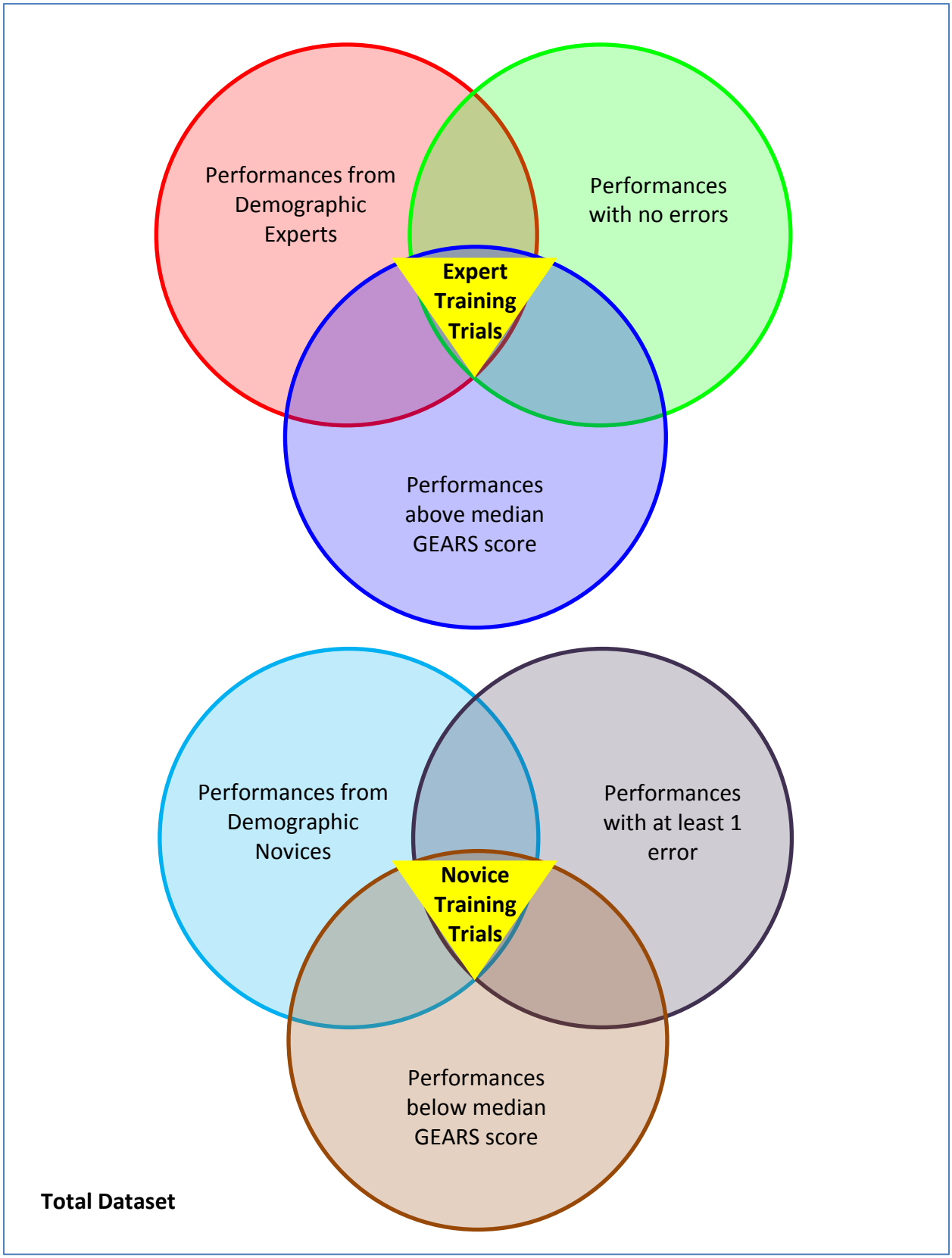


Figure 21 – Criteria for selecting performances used in training skill models.

Within the dataset, 'experts' were defined to include any surgeon who had completed at least 10 cases as the primary surgeon on an actual patient using the da Vinci and at least 10 cases as primary surgeon using laparoscopic tools. All those who did not meet this criterion were to be considered 'novice' for our purposes. A total of 17 of the 51 subjects qualified as demographic experts and 34 qualified as novices. We applied these criteria to the rocking pegboard and suturing performance recordings. The number of performances used to train expert and novice models is listed in Table 17.

Table 17 – Numbers of runs used in training expert and novices models.

	Group	Runs
Rocking Pegboard	Expert Training Trials	9
	Novice Training Trials	10
Suturing	Expert Training Trials	9
	Novice Training Trials	8

4.2.7 Computational Tools

To facilitate fast computational analysis including k-means clustering and HMM training we built a parallelized distributed computing system. Because the SurgTrak codebase is written in Matlab and to promote fast development, we elected to use Matlab Distributed Computing Server (MDCS) to enable distributed computing. MDCS was installed on 10 PCs running Windows 7. The individual Worker Node PCs were networked together and connected to a Head Node to manage and distribute computational tasks. On each of the 10 Worker Node PCs, between two and four worker processes were initiated. The number depends on the PC processor and usually matches the number of CPU cores so that 100% CPU utilization is achieved. Worker Node PC performance specifications varied. The job scheduler used by the MDCS is not sophisticated and does not balance load to higher performance nodes in the

network. The system as a whole was very susceptible to network glitches and if a single node lost connection with the Head Node the entire job would fail. An instance of Matlab running on a Client Node submits tasks to the Head Node and receives results once computation is complete.

In addition to training on the local cloud, models were also trained on a high-end desktop PC for the sake of comparison. The desktop PC featured 24GB of DDR3 RAM, a 10K RPM hard drive and a 2.53 GHz Intel Quad Core Xeon CPU.

4.3 Results and Discussion

4.3.1 Model Training Performance

Table 18 shows the relative gain in performance for model training on the MDCS cluster vs. the high-specification PC. Although the distributed computing approach with MDCS did garner faster training times the difficulty of maintaining one's own network of computers was decidedly not worthwhile. Future researchers should focus on using open source free software and larger scale remote computing resources rather than take this approach.

Table 18 – Relative performance of MDCS vs. robust PC training task execution time.

	MDCS Training Time (Seconds)	Robust PC Training Time (Seconds)	Increase Factor
RPB Expert Model	4560	17967	x 3.94
RPB Novice Model	7417	24982	x 3.36
Suturing Expert Model	1661	5799	x 3.49
Suturing Novice Model	3792	10927	x 2.88

4.3.2 Model Validation

Four models were eventually trained and used for analysis, expert and novice models for rocking pegboard and suturing tasks. Each performance was scored against the expert and novice models for their task. From these raw scores Expert Similarity Factors were computed. These ESF scores and thus the models were validated by comparing scores for data used to train the expert models to data used to train the novice models. For rocking pegboard all but two of the rocking pegboard expert training trials received scores of 0.96 or less, while all but 1 of the rocking pegboard novice training trials scored above 0.96. For the suturing models an ESF score of 0.98 fully discriminated between novice training trials and expert training trials. Thus the models did find performance sequence characteristics in the training data to discriminate between the performances.

Curiously, when the ESF score was compared to a variety of other measures such as the basic measures of Chapter 2, including task time, path length, economy of motion, or such other measures as GEARS score (Chapter 3) or C-SATS score (Chapter 5), essentially no correlative relationship could be identified.

4.3.2 Influence of Warm-Up on Robotic Surgery GEARS Scores

Table 19 shows the mean ESF scores and t-test statistics for each group of performances. We were unable to identify any improvement in performance due to warm-up. Figure 22, Figure 23, and Figure 24 depict boxplots of the ESF scores comparing subjects in the warm-up group with those in the control group for the two tasks and three groupings.

Table 19 –Warm-up impact on HMM ESF scores.

Group	Task	Warm-up	Control	t-test
All Subjects	Rocking Pegboard	0.99	0.98	0.343
Experts	Rocking Pegboard	0.98	0.98	0.716
Novices	Rocking Pegboard	0.99	0.99	0.354
All Subjects	Suturing	0.96	0.96	0.379
Experts	Suturing	0.96	0.96	0.272
Novices	Suturing	0.96	0.96	0.565

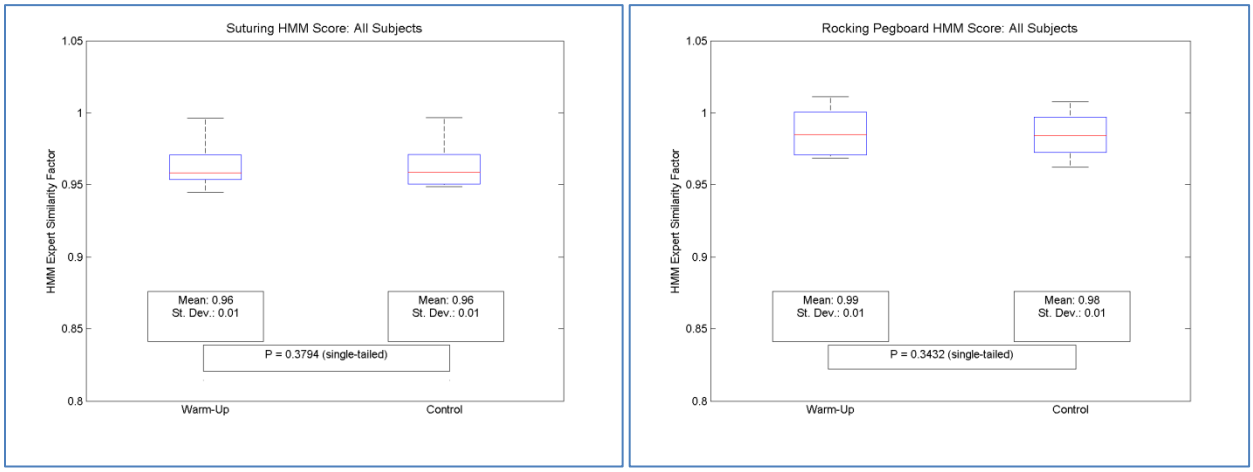


Figure 22 – HMM-based ESF scores for all subjects. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did not generate a noticeable difference in performance.

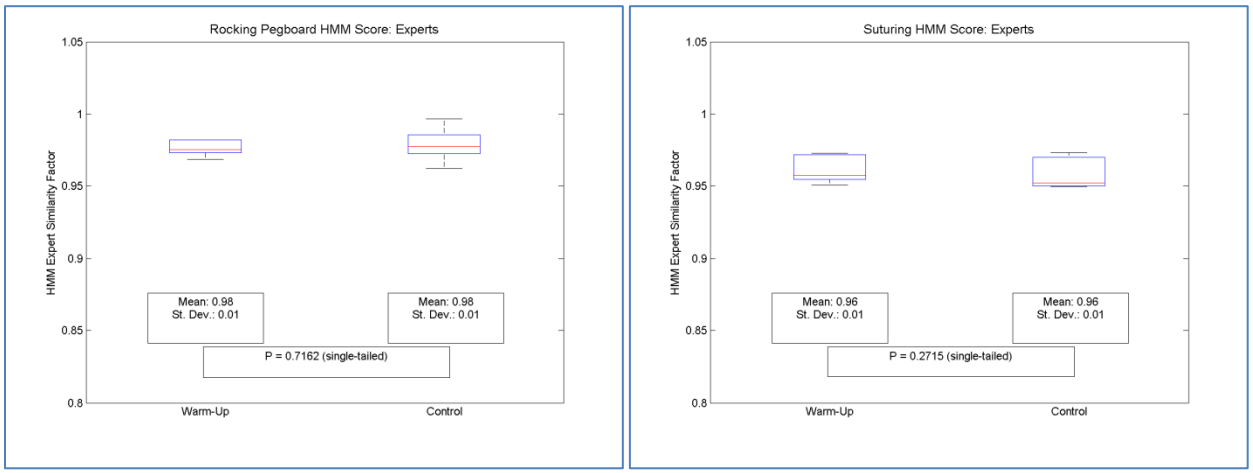


Figure 23 - HMM-based ESF scores for Experts. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did not generate a measurable difference in performance.

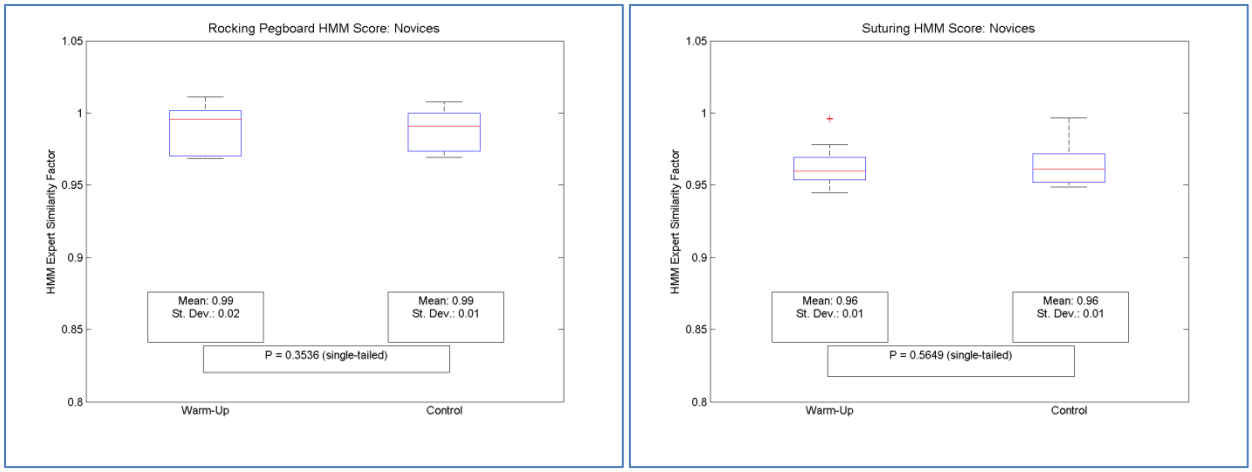


Figure 24 - HMM-based ESF scores for Novices. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did not generate a measurable difference in performance.

4.4 Conclusions

The HMMs developed for this analysis did produce scores which discriminate between the groups of data used to train them. Yet, the ESF scores did not correlate with any other validated measure of performance. This leads us to question the utility of this technique. There may be improved experimental models that would have yielded better results. There also may be characteristics of the data we used that limit the accuracy or utility of the ESF score. In previous successful HMM scoring work force application on the surgical field provided additional data streams to the models [24, 103]. Here we have only motion and velocity data. Also, in previous work, longer surgical procedures were recorded. These procedures had a wider variety of tasks. Among the characteristics that discriminated between the subjects in these studies was time spent idle while novices decided what surgical step to take next. All of the subjects in our study had reached task proficiency. The effects of warm-up thus may not be amenable to analysis using black box HMMs.

Chapter 5: Application of a Novel Crowd-Sourcing Tool to Assess Surgical Performance Following Preoperative Warm-Up

Evaluate the hypothesis that preoperative VR warm-up produces a significant improvement in robotic surgical performance as measured by C-SATS.

5.1 Introduction

Over the years that the BioRobotics Laboratory has been interested in surgical skill evaluation and performance assessment we have come to understand the benefits and drawbacks of the various techniques. Basic measures such as path length, task time and economy of motion are simple to compute and provide immediate results but the resulting measures are not clearly correlates of clinical outcome. Furthermore, the equipment to perform motion tracking can be quite expensive which limit their wider clinical adoption. Machine learning-based algorithms could provide more sensitive data about performance and similarity of an examined performance to known expert performances. However, the methods for successful application are far from established, often requiring hand-tuning of model parameters, and require expensive computational and motion tracking equipment. Not to mention that many of the most successful approaches require force data. Structured assessment tools like GEARS and OSATS seem to be more sensitive to the interaction of the surgeon and the environment, whereas basic measures and algorithms only view the actions of the surgeon without the surgical context. The drawbacks of these assessment tools lie in the high cost of expert surgeon time which leads to a long delay between performance and score.

We seek a method to assess surgical performance which is clinically valid, inexpensive to produce and which can provide assessments in a short period of time. In this chapter I will describe a novel method for assessing surgical performance called Crowd-Sourced Assessment of Technical Skill (C-SATS). This method satisfies all of these criteria. I will describe a pilot study to compare an assessment of a surgical performance performed by surgeons to one performed by a crowd of individuals on the Internet who do not have specific training in surgical skill evaluation. Then I will demonstrate that when applied to a range of performance videos exhibiting a wide variety of skill levels, the crowd agrees well with trained surgeon graders. Finally, I will use C-SATS to measure the impact of VR warm-up on the performance of robotic surgery.

5.1.1 C-SATS Pilot Study

During the summer of 2012 a pilot study was completed to examine the ability of a crowd of untrained individuals to assess a video of a surgical performance [104]. The crowd consisted of workers on an online crowd-sourcing website called Mechanical Turk. Mechanical Turk serves as an Internet marketplace for work which is best done by people, rather than computers. Workers there complete *Human Intelligence Tasks* (HITs) such as completing surveys and categorizing images. Workers are paid for each HIT they complete successfully. A simplified version of the GEARS tool was used consisting of only the Depth Perception, Bimanual Dexterity and Efficiency categories, shown in Figure 25.

Depth perception

1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target

Bimanual dexterity

1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure

Efficiency

1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression

Figure 25 - C-SATS assessment domains.

An online survey was created which first presented a screening video of two surgeons operating side by side on a block transfer task. The subjects were instructed to select the video from the performance they thought to be better in order to test their ability to discriminate between skill levels. Next the subjects read a paragraph of text that instructed them not to answer the question that followed. Both of these attention check questions were intended to eliminate responses from subjects who were not paying attention. Finally, the subjects were presented with a video of a surgeon performing an FLS intracorporeal suturing task using the da Vinci and asked to grade the video using the three categories discussed above. A screenshot from that video is shown in Figure 26.

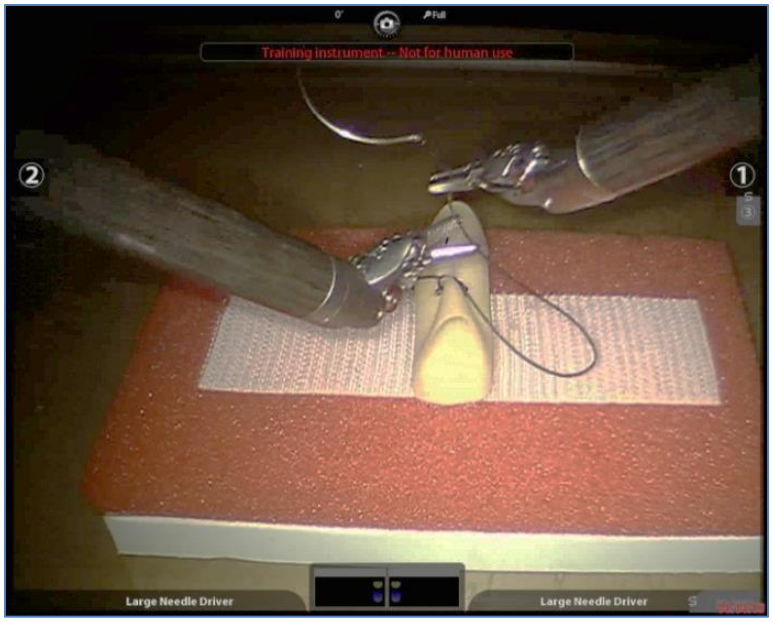


Figure 26 - Criterion video of a surgeon performing an FLS intracorporeal suturing task using the da Vinci surgical robot.

The online survey was sent to 10 recruited faculty-level surgeons. These surgeons were required to have a minimum of three years of minimally invasive surgery experience. The survey was also posted to Facebook to collect voluntary responses. Finally a HIT was created using Mechanical Turk to collect 500 responses from Mechanical Turk workers. Each worker was paid $1.00 (USD) to complete the survey. The score response density for the three groups of responses is shown in Figure 27. Responses from people who failed the attention checking questions were eliminated. This yielded 409 valid Mechanical Turk responses, 9 valid surgeon responses and 67 valid Facebook responses as summarized in Table 20.

Table 20 - Response yield by group.

Group	Total Responses	Correct Responses	Yield %
Mechanical Turk	500	409	81.8 %
Surgeons	10	9	90.0 %
Facebook	110	67	60.9 %

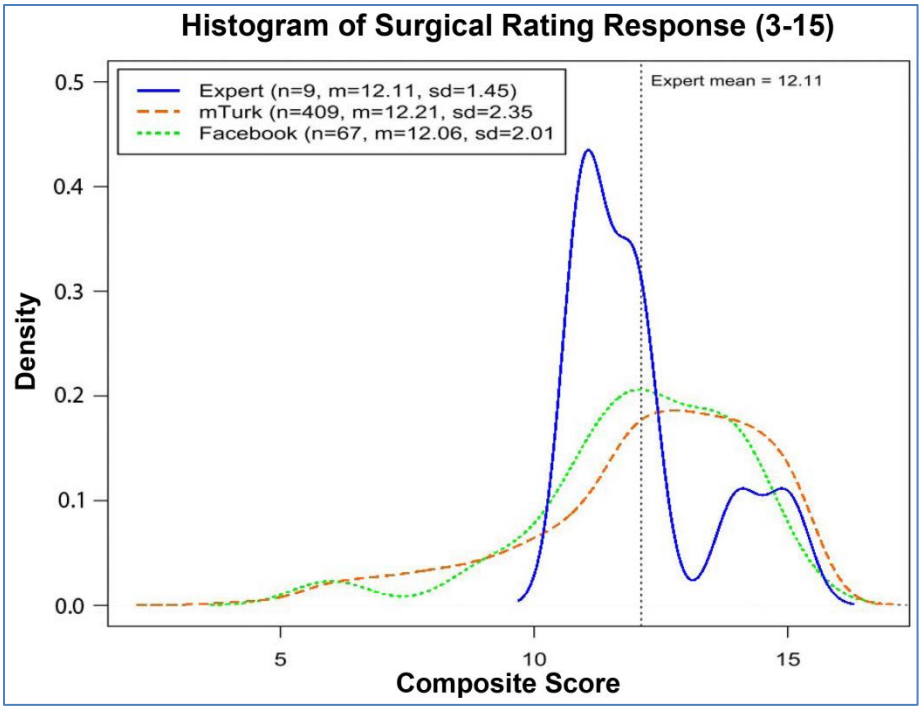


Figure 27 - Scoring density for three groups of scorers: experts, Mechanical Turk workers, and Facebook responses.

Our equivalence criterion was whether the 95% confidence interval of the responses from the Mechanical Turk workers fell entirely within one point of the mean surgeon score, 11.11 to 13.11. The 95% CI of the Mechanical Turk Responses was found to be 11.98 to 12.43, thus satisfying our equivalence criterion. Another most striking result of the pilot study was the response time from the various groups. Shown in table Table 21 and Figure 28, the Mechanical Turker workers were much, much faster to respond.

Table 21 - Time to collect full responses from each group of graders. (* Surveys were released to Mechanical Turk over the course of 5 days but each time they were released the surveys were almost immediately completed by workers. Actual cumulative time to completion disregarding delays between group releases was less than 1 day.)

Group	Days
Mechanical Turk Workers	5*
Surgeons	25
Facebook	24

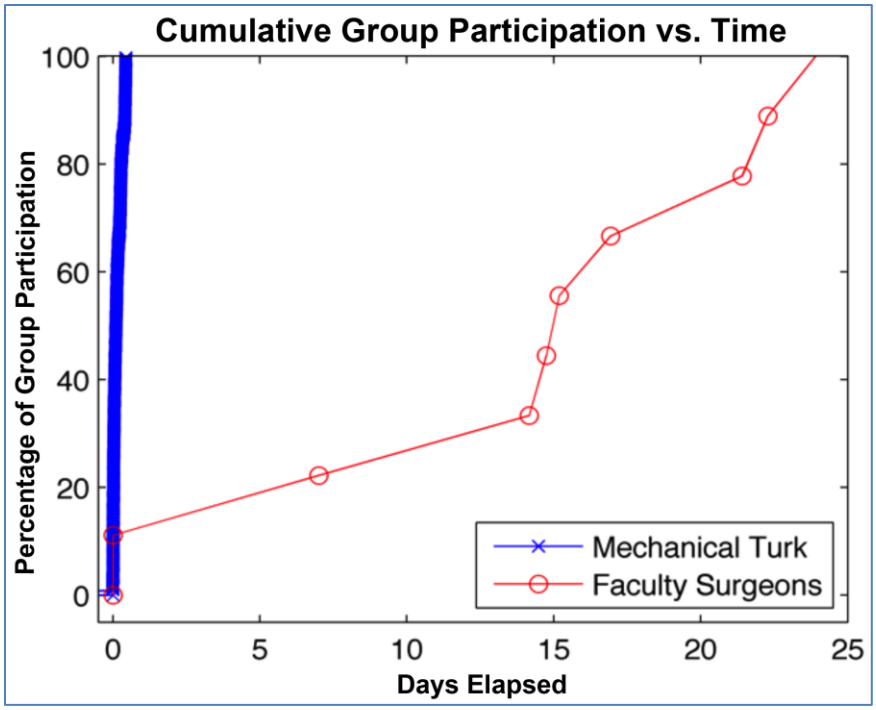


Figure 28 - Simulated completion time for the surgeons vs. the Mechanical Turk workers. The delays between survey releases on Mechanical Turk have been eliminated.

This pilot study satisfied us that Mechanical Turk workers could grade comparably to surgeon graders and that wider examinations of these phenomena were needed. Furthermore, it showed that the Mechanical Turk workers were willing to provide responses much more quickly than surgeons and for a potentially much lower cost.

5.2 Methods

5.2.1 Data

The same set of tasks scored using GEARS by the trained surgeon graders, discussed in Chapter 3, were selected to be scored using C-SATS. This allows us to compare the crowd scores to the surgeon scores. Selecting one rocking pegboard task and one suturing task allowed us to comment on the impact of warm-up when the warm-up VR task was either similar to or

different from the criterion task. 51 subjects proceeded through the study. This should have yielded 51 videos for both the rocking pegboard task and the suturing task. However, 2 of the 51 performances for the rocking pegboard task and 2 of the 51 performances for the suturing task were not recorded due to errors in the recording equipment. Thus, in both cases, 49 videos were available for analysis. Videos from session 3 of the rocking pegboard task were selected as these performances occurred the longest time from the proficiency phase of the study. This should minimize the influence of having had recent significant practice on the task.

Table 22 - Tasks scored by expert surgeons using C-SATS were the same as those scored using GEARS.

Performance Type	Performances Scored	Session Number
Rocking Pegboard	49	3
Suturing	49	4

5.2.2 C-SATS Mechanical Turk Crowd Survey

We adapted the HTML Form-based GEARS Grading Suite to create a survey similar to that used in the pilot study. The survey including screening questions followed by the three grading domains with free text response fields is shown in Figure 29. Surveys for each task to be graded were automatically generated using a Matlab script. A PHP CGI script on a BioRobotics Lab web server received the survey responses, stored the scores to our server and generated a unique survey code that the workers copied into the Mechanical Turk website. We created a HIT requesting 30 crowd responses for each of the 49 rocking pegboard and 49 suturing tasks using the Mechanical Turk web interface. We decided to collect 30 responses from the crowd because we believed this to be a number sufficient to judge the overall agreement between surgeons and the crowd and a sufficiently high number to get a sample mean that should be representative of the crowd response population mean. Furthermore, this was a number that

we could afford to request as we had to pay for each response generated. Table 23 describes the HIT parameters including the pay per tasks completed. Mechanical Turk manages the assignment of HITS to workers so that each worker may complete multiple HITS but they may only complete a given HIT once. Thus the 30 responses collected per performance are from unique workers. Also, since some of the workers will answer the attention check questions incorrectly, the work from these workers was rejected and the HIT relaunched for other workers to complete. This allowed us to assure we collected at least 30 valid responses per performance.

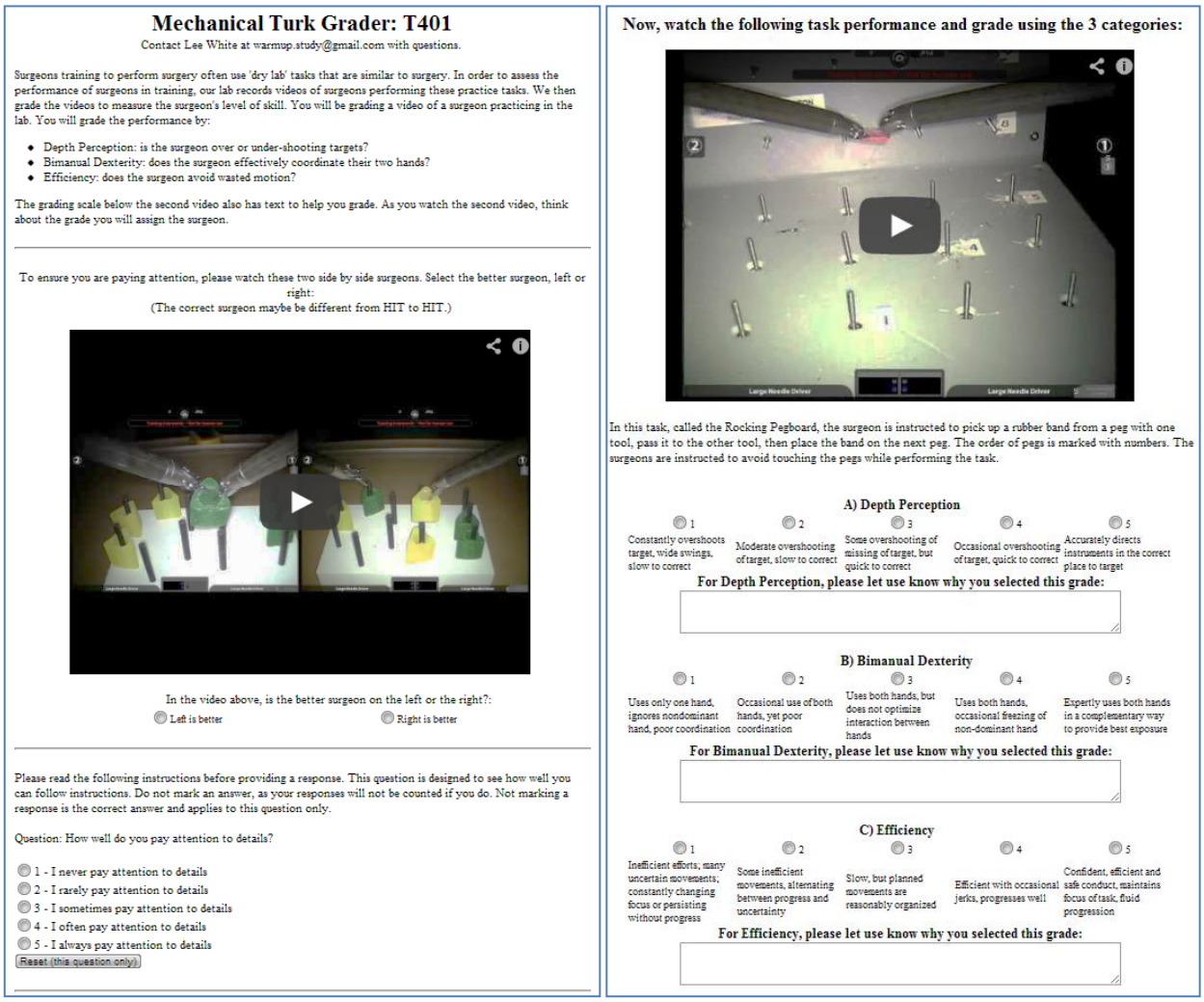


Figure 29 - C-SATS survey for Mechanical Turk workers. LEFT: screening questions. RIGHT: grading domains with free text response areas.

Table 23 – C-SATS survey characteristics.

	Rocking Pegboard	Suturing
Pay	$0.25	$0.50
HITs per performance	30	30

To provide additional insight into general trends in the accuracy of how crowd-derived performance scores varied across the spectrum of performance levels, that is, to answer the question "does the crowd score poor performances as accurately as excellent performances?" three additional performances from each task were each presented to 150 workers. The selected tasks were at the 10^{th}, 50^{th}, and 90^{th} percentile level within the range of performances according to surgeon-derived C-SATS score. This allowed us to analyze the performance quality-dependent bias in the crowd.

Table 24 – C-SATS deep bias survey characteristics.

	Rocking Pegboard	Suturing
Pay	$0.50	$0.50
HITs per performance	150	150

5.3 Results and Discussion

5.3.1 Response Statistics

We launched the rocking pegboard task first and paid $0.25 per completed HIT. This seemed to be a little low for the amount of time needed to complete the task. As can be seen in Table 25, the completion time was must faster for the following task, suturing, where we paid $0.50. Another observation about the responses we collected was that perhaps because of price sensitivity and because the HIT was launched in the morning in the United States, the majority of responses were from the US (based on worker self-reporting). See Figure 30 and Figure 31.

Table 25 – C-SATS survey response characteristics.

	Rocking Pegboard	Suturing
Pay	$0.25	$0.50
Total Responses	2027	1668
Valid Responses	1433	1498
Yield %	70.7%	89.8%
Cost	$493.75	$768.00
Completion Time (95%)	108:48:00	8:52:00
Domestic Responses	37.79%	94.60%

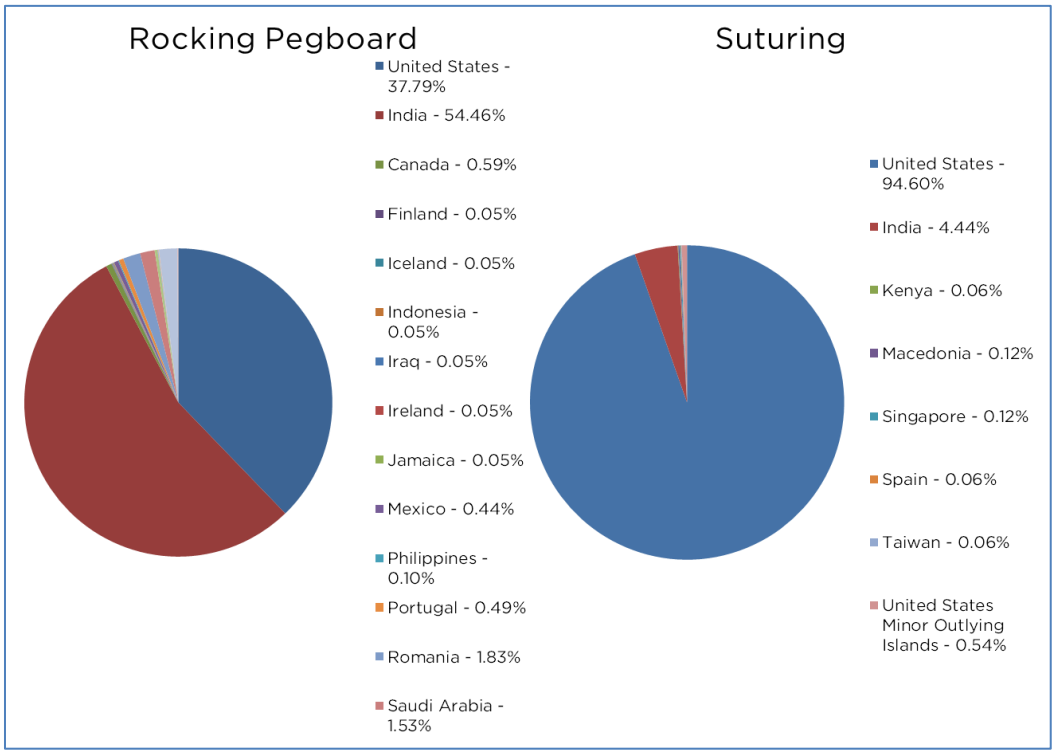


Figure 30 - Crowd worker self-reported location.

We also observed a much higher yield (correct responses out of total responses) for the suturing task. Based on the predominance of domestic responses for the suturing task, one might suspect the higher yield to be due to the English comprehension capabilities of the workers. Further analysis might indicate that domestic workers produce a higher proportion of valid responses.

Figure 31 – Locations of Mechanical Turk Workers. GREEN: Self-reported locations. RED: IP-derived locations for the Rocking Pegboard task. BLUE: IP-Derived locations for the Suturing task.

5.3.2 Validating Crowd-Sourced Assessment of Technical Skills (C-SATS) using Amazon Mechanical Turk

To compare the grades provided by the surgeon graders to those provided by the crowd, simulated surgeon C-SATS scores were generated from the 5 domain scores collected in the GEARS grading portion of the study. The scores provided by the surgeons in the 3 gears domains of C-SATS were isolated and summed to produce a surgeon C-SATS score. The surgeons were not required to answer the attention check questions. The simulated surgeon C-SATS score was then compared to the score provided by the members of the crowd who

answered the screening questions correctly. Figure 32, Figure 33 and Figure 34 show the relationship between the surgeon C-SATS scores and the crowd C-SATS scores. The correlation coefficient between surgeon score and crowd score was found to be 0.79 for the rocking pegboard task and 0.86 for the suturing task, indicating the scores were very highly correlated. The equivalence criteria developed in the pilot study showed that for that video 95% of the time the crowd score was within 1 point of the surgeon score. If that were to occur in this experiment we would expect almost all of the crowd scores to fall within the dotted lines on their side of the lines of slope 1 in Figure 33. In our collected data across a wider range of performance levels it was seen that 16 and 15 scores for rocking pegboard and suturing, respectively, fell outside this equivalence zone.

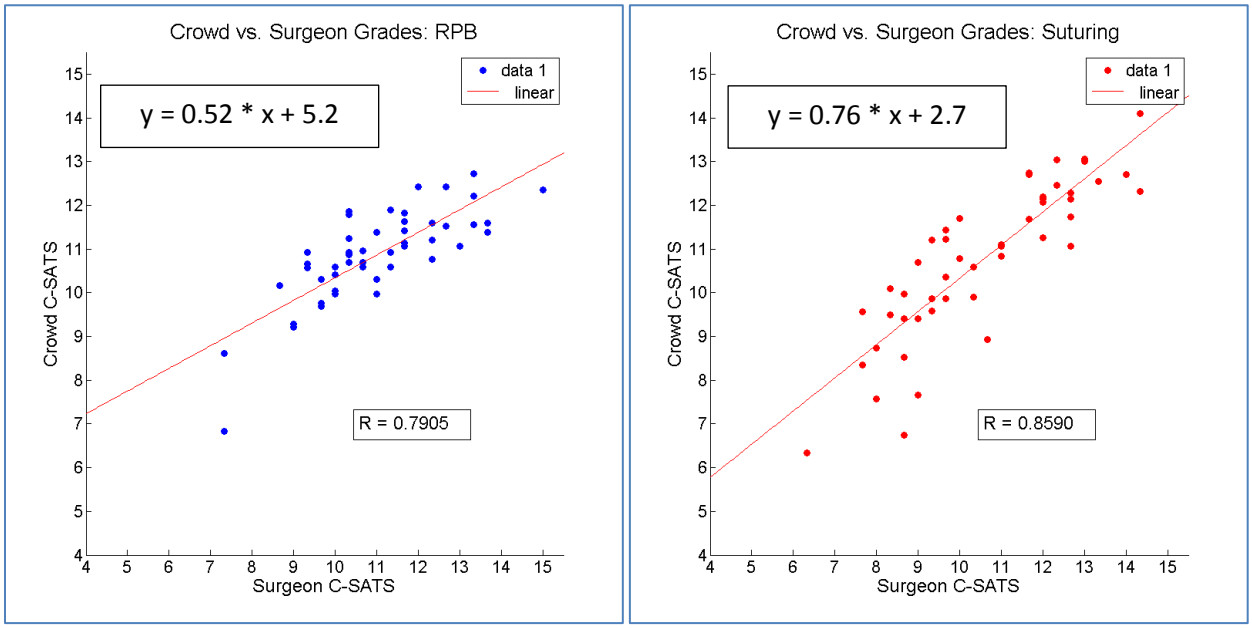


Figure 32- Crowd-Surgeon correlation coefficients and lines of best fit. LEFT: Rocking Pegboard. RIGHT: Suturing.

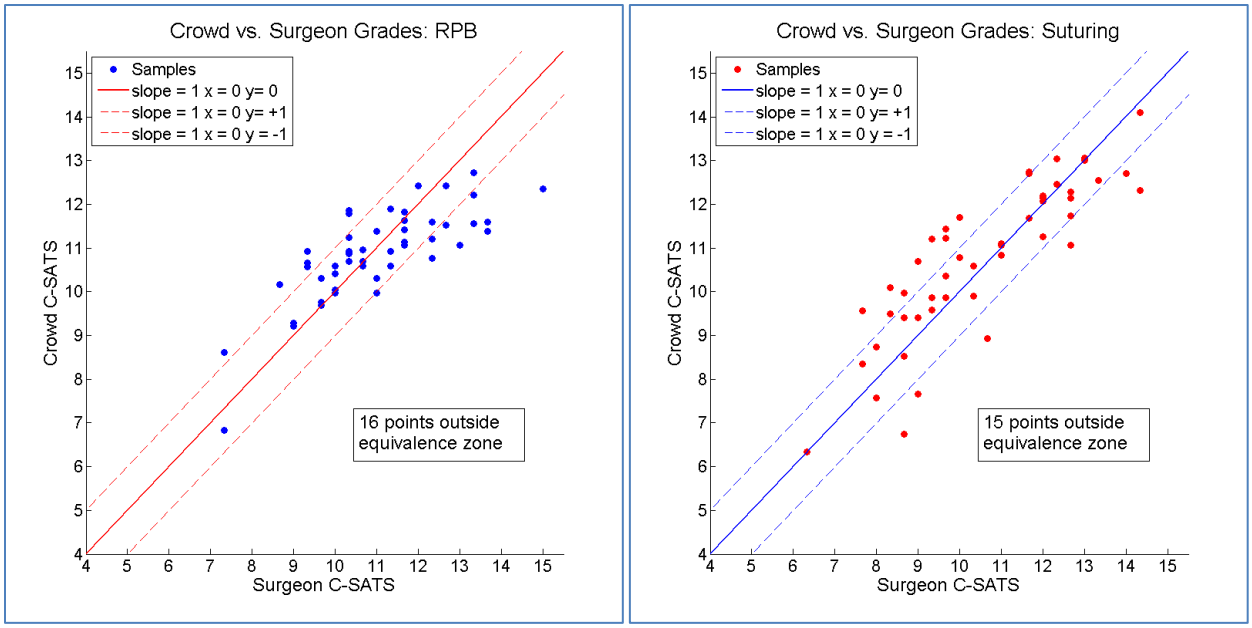


Figure 33 - Crowd-Surgeon equivalence analysis. LEFT: Rocking Pegboard. RIGHT: Suturing. If the crowd grade is within 1 point of the surgeon grade the point will fall within the dotted lines.

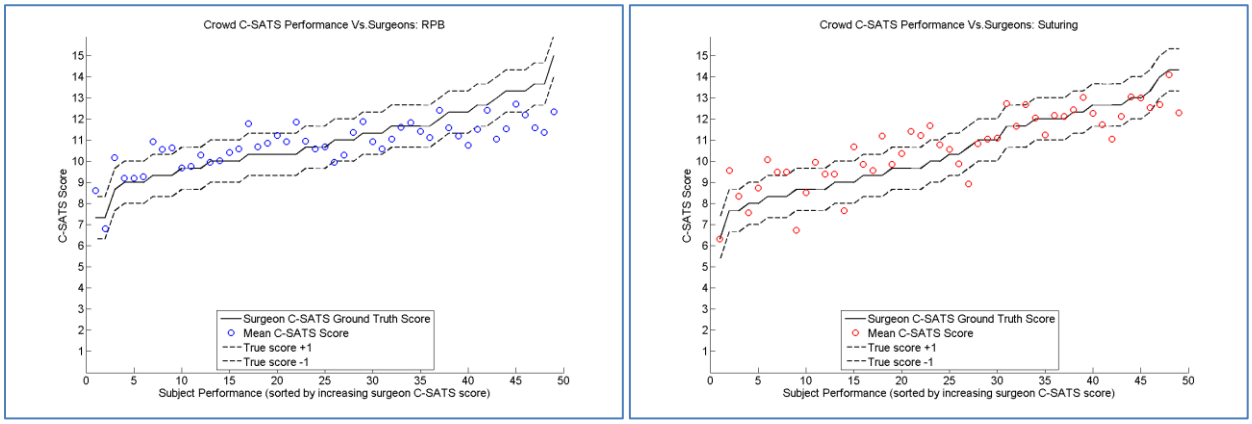


Figure 34 – Another way to depict the agreement between the workers and the surgeons. LEFT: Rocking Pegboard. RIGHT: suturing. It can be seen that the workers seem not to grade the best performances as highly for the rocking pegboard task.

It can also be observed that for the rocking pegboard task the crowd scored more critically those performances which the surgeons scored at the higher end of the performance spectrum. Results from the deep bias analysis in which three performances from both tasks were presented to 150 members of the crowd showed the same bias for better rocking pegboard performances. See Figure 34. While the crowd bias was flat across the performance spectrum,

the crowd disagreed with the surgeons for the 90th percentile rocking pegboard performance, scoring nearly two points more critically.

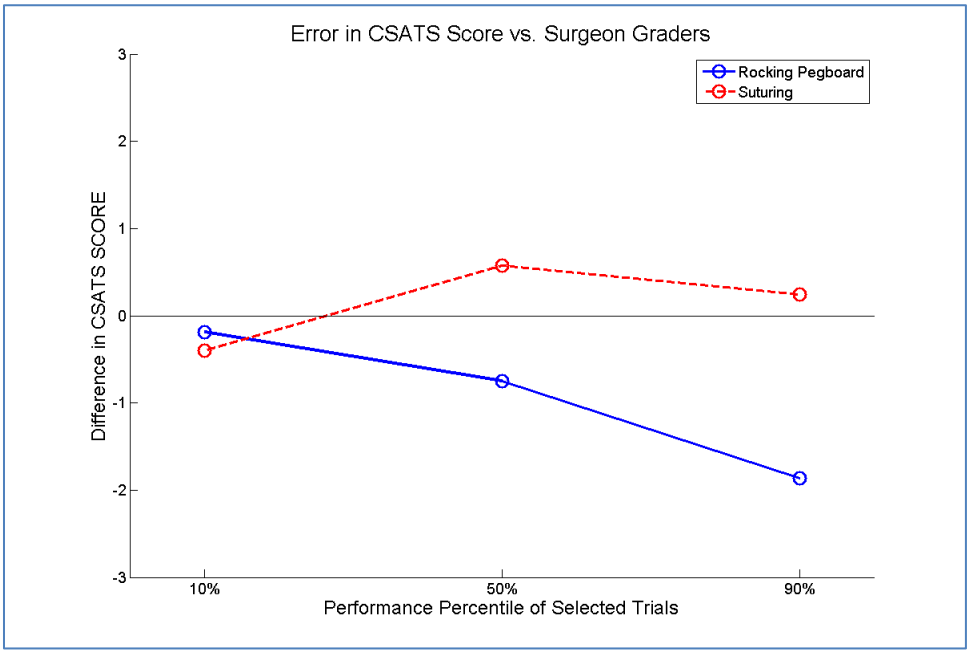


Figure 35 – Deep bias analysis shows correlation between surgeon score and crowd score for 3 performances each from the two tasks. The selected tasks were the 10%, 50%, 90% of the range of performances according to surgeon-derived C-SATS score. Then 150 Mechanical Turk workers were recruited to grade the performances, with the aim being to determine if the accuracy of the workers is dependent on the quality of the performance.

Understanding this task-and-performance-quality-dependent bias could allow for a correction factor to be applied in future efforts, thus recovering a more accurate estimation of a surgeon derived score from the crowd.

Next, we computed the level of agreement between the surgeons as a group and the crowd as a group. Using Cronbach's alpha, and treating the mean surgeon C-SATS score as one grader and the mean crowd C-SATS score as another, the agreement on the rocking pegboard task score 0.84 and the agreement on the suturing task scored 0.92.

Table 26 – Agreement when surgeon-derived C-SATS scores are compared with crowd-derived C-SATS scores.

	Rocking Pegboard	Suturing
Cronbach's alpha	0.84	0.92
Agreement Quality	Good	Excellent

By a variety of metrics we see that the crowd can provide high-quality performance assessment metrics for the assessment of surgical performance. This study focused on the dry-lab setting. Further assessment, refinement and development are needed to see if these strong results hold in the setting of actual surgeries. Further work is also needed to determine the best way to have the crowd score longer videos or segments of tasks longer in duration than those scored here.

5.3.3 Warm-up Results

Having shown the C-SATS score to be an accurate and useful tool to assess surgical performance, we also tested the warm-up hypothesis using the grades from the crowd. When considering the impact of warm-up on rocking pegboard and suturing, we first look at the performances as a group. As seen in Figure 36, for the rocking pegboard task the warm-up group achieved a mean C-SATS score of 10.87 while the control group performed slightly worse, scoring 10.57 on average. For the suturing task the warm-up group outperforms the control group 11.10 to 10.55. In the rocking pegboard case the differences between the distributions of scores did not achieve statistical significance, but in the case of suturing, the warmed up group was statistically significantly better.

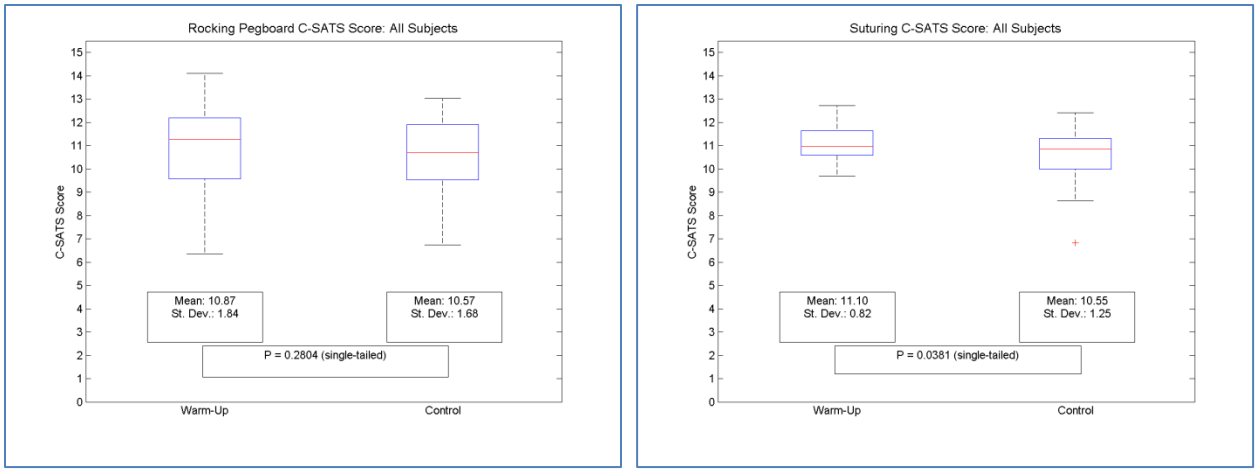


Figure 36 – C-SATS-based scores for all subjects. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did generate a statistically significant difference in performance for the suturing task.

When we separate the groups into experts and novices and consider their performances separately, we find that in all but one case (novices on the rocking pegboard task), C-SATS scores favor warm-up, however only for experts does this difference achieve statistical significance at the level we selected. These results are depicted in Figure 37 and in Figure 38 as well as summarized in Table 27.

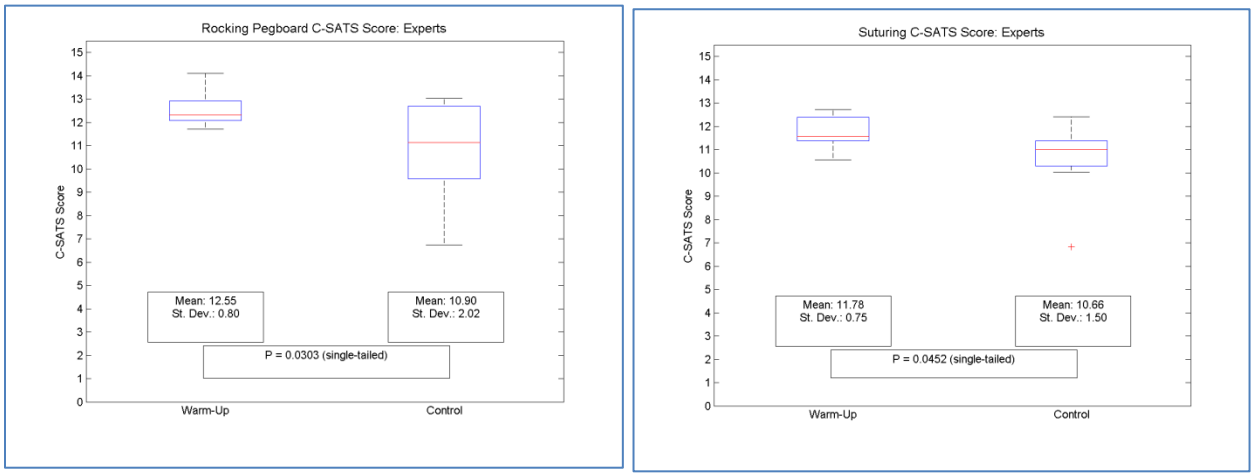


Figure 37 - C-SATS-based scores for experts. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did generate a statistically significant difference in performance for both tasks.

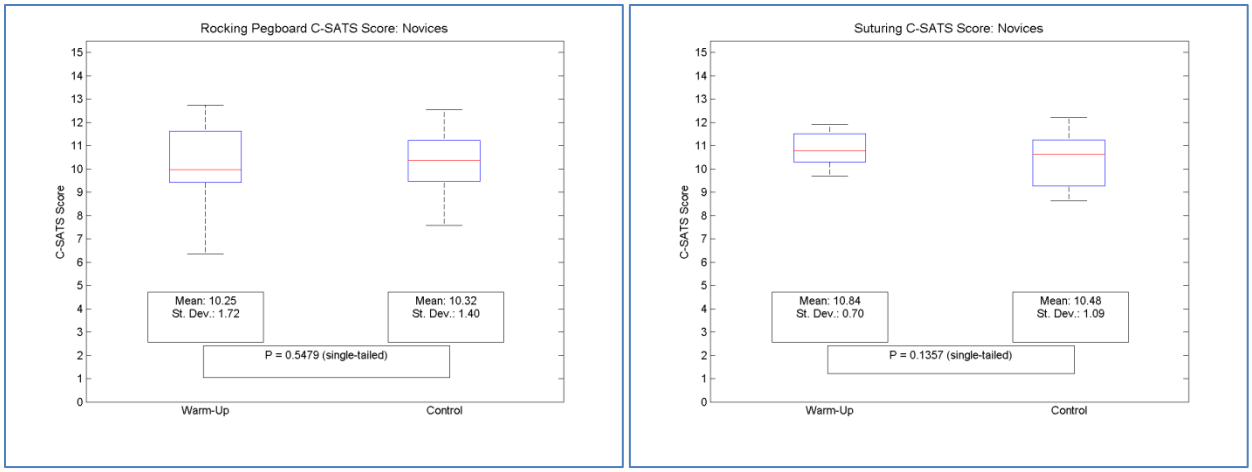


Figure 38 - C-SATS-based scores for novices. LEFT: Rocking Pegboard. RIGHT: Suturing. Warm-up did generate a statistically significant difference in performance for the suturing task.

Table 27 –Warm-up impact on surgeon C-SATS scores. (*statistically significant)

Group	Task	Warm-up	Control	t-test
All Subjects	Rocking Pegboard	10.87	10.57	0.280
Experts	Rocking Pegboard	12.55	10.90	0.030 *
Novices	Rocking Pegboard	10.25	10.32	0.548
All Subjects	Suturing	11.10	10.55	0.038 *
Experts	Suturing	11.78	10.66	0.045 *
Novices	Suturing	10.84	10.48	0.136

These results are very similar to those found in the GEARS score analysis. Here though, the differences between the warm-up group and the control group did achieve statistical significance for three categories: all subjects on the suturing task and experts for both tasks.

5.4 Conclusions

5.4.1 C-SATS Utility

C-SATS allows researchers and educators to grade surgical performance videos much faster than is usually possible using structured assessment tools. The technique produces scores which are highly correlated with assessments by trained surgeons, as measured by correlation coefficient and by Cronbach's alpha. In most cases the scores provided by the crowd are within

1 point of that provided by our group of surgeons. Additional refinement of the application of C-SATS could further improve this agreement. It may be possible to eliminate performance quality bias in the crowd responses. This tool could be very useful for assessing large sets of data where expert surgeon grading would be impractical or impossible. There is a cost to using the tool but it is less than the cost to have surgeons score performances. Furthermore, it is a better use of surgeon resources. Finally, there are communities of people willing to do similar crowd-sourced work for the advancement of science without getting paid. Such a system for assessing surgical performance of trainees may be one way to further lower the cost to perform C-SATS assessments. Because of the existence of the Internet and surgical systems that allow for immediate and possibly automatic upload or real-time streaming of surgical video feeds, a future where crowd-sourced surgical skill assessment is a reality could be imminent.

5.4.2 C-SATS Measurement of Warm-up

The crowd seemed to be more sensitive to the effects of warm-up. They observed a benefit in all but one case (novice surgeons on rocking pegboard) and those benefits where statistically significant in three of those cases. Interestingly, expert surgeons benefitted most from warm-up, both when the VR warm-up was similar to the criterion task and when it was different. The benefit of warm-up in a clinical setting should be measured. Evidence so far indicates that it could help. The monetary and time costs of having surgeons sit down to a simulator, especially when that simulator is actually the master console of the robot itself, could be acceptably low for widespread implementation.

Chapter 6: Summary

The promise of identifying a convenient deployable tool to maximize robotic surgical performance prompted our team to devise and complete this series of studies. Via three of four analysis approaches we have identified clear benefits to warming up before performing robotic surgery. These results are summarized in Table 28. Further examination is needed, including testing the warm-up hypothesis in the clinic. Furthermore, a study structure that directly controls for individual performance will be utilized in all future studies.

Table 28 – Warm-up impact on surgical performance assessed by a variety of metrics and displayed as corresponding test statistics. Green shading: mean favors warm-up but not statistically significant. Red shading: Statistically significant improvement. Rocking pegboard: Error 1 = peg touches, Error 2 = cognitive error. Suturing: Error 1 = technical error, Error 2 = global error score. Cells where a test statistic was not computed are left blank.

	Rocking Pegboard			Suturing		
Metric	All Subjects	Experts	Novices	All Subjects	Experts	Novices
Basic Measures						
EOM	0.132	0.007	0.9	0.557		
Path	0.014	0.093	0.063	0.999		
Time	0.001	0.001	0.10	0.703		
Error 1	0.313	0.60	0.18	0.115		
Error 2	0.340	0.96	0.27	0.014		
GEARS	0.522	0.054	0.613	0.092	0.051	0.228
C-SATS	0.280	0.030	0.548	0.038	0.045	0.136
HMM	0.343	0.716	0.354	0.379	0.272	0.565

Regarding the methods for analyzing surgical performance, each had their pros and cons in terms of convenience, cost, and clinical relevance, as can be seen in Table 29. C-SATS is a new and exciting way to assess surgical performance. It correlates well with the gold standard assessment of performance using expert surgeons. This is clearly an area with great potential for research and clinical application.

Table 29 – Pros and cons of each method for assessing surgical performance.

	Pros	Cons
Basic Measures	• Easy to compute • Potentially OR compatible (especially if fully robot integrated) • Some measures, like task time, linked to clinical outcome	• Tracking technology is expensive • cumulative error tabulation is time consuming for a proctor (but potentially crowd-source-able) • Not clearly linked to surgical outcome • Limited to surgical procedures where motion data is available • Limited surgical field context
GEARS	• Widely accepted • Potentially applicable to a wide variety of type of surgery and other medical interventions • Leverages the perceptive power of the human brain • OR compatible	• Expensive in terms of expert surgeon time • Needs further clinical validation • Long time delay between performance and score coming back • potential for patient privacy issues
C-SATS	• Based on an accepted clinical performance measure • Scalable to large amount of performance data • Potentially applicable to a wide variety of type of surgery and other medical interventions • Short time from performance to score, possibly less than an hour • Leverages the perceptive power of the human brain • OR compatible	• Needs further clinical validation • Potential for patient privacy issues • May require training for improved scoring • Expensive for large datasets • Needs careful control of crowd to avoid bad data and system abuse
HMM	• Essentially zero marginal cost to assess a performance once model is built and refined • Immediate scores for existing skill models • Potentially OR compatible	• Tracking technology is expensive • Skill model technology is still at the research stage • Model training can be expensive and time consuming • No surgical field context • Only aware of surgeon movements

Citations

[1] Gallagher AG, Boyle E, Toner P, Neary PC, Andersen DK, Satava RM, et al. Persistent next-day effects of excessive alcohol consumption on laparoscopic surgical performance. Archives of Surgery. 2011;146(4):419.

[2] Alexander AD. Impacts of telemation on modern society. In: Human Factors and Ergonomics Society Annual Meeting Proceedings. vol. 17. Human Factors and Ergonomics Society; 1973. p. 299–304.

[3] Kwoh YS, Hou J, Jonckheere EA, Hayati S. A robot with improved absolute positioning accuracy for CT guided stereotactic brain surgery. Biomedical Engineering, IEEE Transactions on. 1988;35(2):153–160.

[4] Cuschieri A. The laparoscopic revolution–walk carefully before we run. Journal of the Royal College of Surgeons of Edinburgh. 1989;34(6):295.

[5] Centres R. Cholecystectomy practice transformed. The Lancet. 1991;338(8770):789 – 790. Originally published as Volume 2, Issue 8770. Available from: http://-www.sciencedirect.com/science/article/pii/014067369190672C.

[6] Intuitive Surgical. Investor Presentation Q2 2012; 2012.

[7] Intuitive Surgical. Investor Presentation Q1 2013; 2013. Available from: http://-www.morningstar.com/earnings/earnings-call-transcript.aspx?t=ISRG.

[8] Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ. Effects of limited work hours on surgical training. Journal of the American College of Surgeons. 2002;195(4):531–538.

[9] Gawande A. Two Hundred Years of Surgery. New England Journal of Medicine. 2012;366(18):1716–1723. Available from: http://www.nejm.org/doi/full/10.1056/-NEJMra1202392.

[10] Shreve J, of Actuaries Health Section S, (Firm) M. The economic measurement of medical errors. Society of Actuaries; 2010. Available from: http://soa.org/Files/Research/-Projects/research-econ-measurement.pdf.

[11] Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA: the journal of the American Medical Association. 2003;290(14):1868–1874.

[12] Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. Surgical endoscopy. 2011;25(2):356–366. Available from: https://cirl.lcsr.jhu.edu/-wiki/images/5/5d/SurgicalEndoscopy2008_rgr_final.pdf.

[13] Murphy SL, Xu J, Kochanek KD. Deaths: Preliminary Data for 2010. National Vital Statistics Reports. 2012;60(4). Available from: http://www.cdc.gov/nchs/data/nvsr/nvsr60/-nvsr60_04.pdf.

[14] Commission MQA. Continuing Education Requirements Frequently Asked Questions for Physicians; 2012. Online. Available from: http://doh.wa.gov/hsqa/MQAC/-PhysicianEdu.htm#HowOften.

[15] ABS to Require ACLS, ATLS and FLS for General Surgery Certification; 2008. Online. Available from: http://www.absurgery.org/default.jsp?news_newreqs.

[16] Page L. Robot Maker Sued Over Hysterectomy Patient's Death; 2012. Online. Available from: http://www.outpatientsurgery.net/news/2012/04/6-Robot-Maker-Sued-Over-Hysterectomy-Patient-s-Death.
[17] Ostrom CM. Failed robotic surgery focus of Kitsap trial. The Seattle Times. 2013;Available from: http://seattletimes.com/html/localnews/2020918732_robottrialxml.html.
[18] Anonymous. Psychomotor learning; 2012. Online. Available from: http://-en.wikipedia.org/wiki/Psychomotor_learning.
[19] Sturm LP, Windsor JA, Cosman PH, Cregan P, Hewett PJ, Maddern GJ. A systematic review of skills transfer after surgical simulation training. Annals of surgery. 2008;248(2):166.
[20] Rosen J, Solazzo M, Hannaford B, Sinanan M. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Computer Aided Surgery. 2002;7(1):49–61.
[21] Satava RM, Cuschieri A, Hamdorf J. Metrics for objective assessment. Surgical endoscopy. 2003;17(2):220–226.
[22] Carter BN. The fruition of Halsted's concept of surgical training. Surgery. 1952;32(3):518.
[23] Vemer L, Oleynikov D, Holtmann S, Haider H, Zhukov L. Measurements of the Level of Surgical Expertise Using Flight Path Analysis from da Vinci™ Robotic Surgical System. Medicine meets virtual reality 11: NextMed: health horizon. 2003;94:373.
[24] Rosen J, Brown JD, Chang L, Sinanan M, Hannaford B. Generalized Approach for Modeling Minimally Invasive Surgery as a Stochastic Process Using a Discrete Markov Model. IEEE Transactions on Biomedical Engineering. 2006 Mar;53(3):399–413.
[25] Jaffer AK, Barsoum WK, Krebs V, Hurbanek JG, Morra N, Brotman DJ. Duration of anesthesia and venous thromboembolism after hip and knee arthroplasty. In: Mayo Clinic Proceedings. vol. 80. Mayo Clinic; 2005. p. 732–738.
[26] Ferrier MB, Spuesens EB, Le Cessie S, Baatenburg de Jong RJ. Comorbidity as a major risk factor for mortality and complications in head and neck surgery. Archives of Otolaryngology-Head and Neck Surgery. 2005;131(1):27.
[27] Martin J, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery. 1997;84(2):273–278.
[28] Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global Evaluative Assessment of Robotic Skills: Validation of a Clinical Assessment Tool to Measure Robotic Surgical Skills. The Journal of Urology. 2011;187:247–252.
[29] van Hove P, Tuijthof G, Verdaasdonk E, Stassen L, Dankelman J. Objective assessment of technical surgical skills. British Journal of Surgery. 2010;97(7):972–987.
[30] Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological bulletin. 1955;52(4):281.
[31] Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, et al. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. Journal of Laparoendoscopic & Advanced Surgical Techniques. 2000;10(4):183–190.
[32] Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. The American journal of surgery. 2006;192(3):372–378.

[33] Rosen J, Solazzo M, Hannaford B, Sinanan M. Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions. Studies in health technology and informatics. 2001;p. 417–423.
[34] Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. Biomedical Engineering, IEEE Transactions on. 2002;48(5):579–591.
[35] Lin H, Shafran I, Murphy T, Okamura A, Yuh D, Hager G. Automatic detection and segmentation of robot-assisted surgical motions. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005. 2005;p. 802–810.
[36] Leong J, Nicolaou M, Atallah L, Mylonas G, Darzi A, Yang GZ. HMM assessment of quality of movement trajectory in laparoscopic surgery. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006. 2006;p. 752–759.
[37] Reiley C, Hager G. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009. 2009;p. 435–442.
[38] Brown JD. In-Vivo and Postmortem Biomechanics of Abdominal Organs Under Compressive Loads: Experimental Approach in a Laparoscopic Surgery Setup. PhD Thesis. 2003 Dec;.
[39] Richards C, Rosen J, Hannaford B, Pellegrini C, Sinanan M. Skills evaluation in minimally invasive surgery using force/torque signatures. Surgical Endoscopy. 2000;14(9):791–798.
[40] Kowalewski TM, Rosen J, Chang L, Sinanan M, Hannaford B. Optimization of a vector quantization codebook for objective evaluation of surgical skill. In: Proc. Medicine Meets Virtual Reality 12; 2004. p. 174–179.
[41] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.
[42] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
[43] Reiley CE, Lin HC, Varadarajan B, Vagvolgyi B, Khudanpur S, Yuh DD, et al. Automatic recognition of surgical motions using statistical modeling for capturing variability. Studies in health technology and informatics. 2008;132:396.
[44] Volianitis S, McConnell AK, Koutedakis Y, Jones DA. Specific respiratory warm-up improves rowing performance and exertional dyspnea. Medicine & Science in Sports & Exercise. 2001;33(7):1189.
[45] Knudson DV, Noffal GJ, Bahamonde RE, Bauer JA, Blackwell JR, et al. Stretching has no effect on tennis serve performance. Journal of strength and conditioning research/National Strength & Conditioning Association. 2004;18(3):654.
[46] Edwards B, Edwards W, Waterhouse J, Atkinson G, Reilly T, et al. Can cycling performance in an early morning, laboratory-based cycle time-trial be improved by morning exercise the day before? International journal of sports medicine. 2005;26(8):651–656.
[47] Guidetti L, Emerenziani GP, Gallotta MC, Baldari C. Effect of warm up on energy cost and energy sources of a ballet dance exercise. European journal of applied physiology. 2007;99(3):275–281.

[48] Small K, Mc Naughton L, Matthews M. A systematic review into the efficacy of static stretching as part of a warm-up for the prevention of exercise-related injury. Research in Sports Medicine. 2008;16(3):213–231.
[49] Hajoglou A, Foster C, de Koning JOSJ, Lucia A, Kernozek TW, Porcari JP. Effect of warm-up on cycle time trial performance. Medicine and Science in Sports and Exercise. 2005;37(9):1608.
[50] Bishop D. Warm up I: potential mechanisms and the effects of passive warm up on exercise performance. Sports Medicine. 2003;33(6):439–454.
[51] Anshel MH, Wrisberg CA. Reducing warm-up decrement in the performance of the tennis serve. Journal of Sport & Exercise Psychology. 1993;.
[52] Schmidt RA, Lee TD. Motor control and learning: A behavioral emphasis. Human Kinetics Publishers; 2005.
[53] Do AT, Cabbad MF, Kerr A, Serur E, Robertazzi RR, Stankovic MR. A warm-up laparoscopic exercise improves the subsequent laparoscopic performance of Ob-Gyn residents: a low-cost laparoscopic trainer. JSLS: Journal of the Society of Laparoendoscopic Surgeons. 2006;10(3):297.
[54] Kahol K, Satava RM, Ferrara J, Smith ML. Effect of Short-Term Pretrial Practice on Surgical Proficiency in Simulated Environments: A Randomized Trial of the “Preoperative Warm-Up” Effect. Journal of the American College of Surgeons. 2009;208(2):255–268.
[55] Kahol K, Krishnan NC, Balasubramanian VN, Panchanathan S, Smith M, Ferrara J. Measuring movement expertise in surgical tasks. In: Proceedings of the 14th annual ACM international conference on Multimedia. ACM; 2006. p. 719–722.
[56] Calatayud D, Arora S, Aggarwal R, Kruglikova I, Schulze S, Funch-Jensen P, et al. Warm-up in a virtual reality environment improves performance in the operating room. Annals of surgery. 2010;251(6):1181.
[57] Gallagher A, McClure N, McGuigan J, Ritchie K, Sheehy N, et al. An ergonomic analysis of the fulcrum effect in the acquisition of endoscopic skills. Endoscopy. 1998;30:617–620.
[58] Quinn AJ, Bederson BB; ACM. Human computation: a survey and taxonomy of a growing field. 2011;p. 1403–1412.
[59] Zaidan OF, Callison-Burch C. Crowdsourcing translation: Professional quality from non-professionals. 2011;1:1220–1229.
[60] Noronha J, Hysen E, Zhang H, Gajos KZ; ACM. Platemate: crowdsourcing nutritional analysis from food photographs. 2011;p. 1–12.
[61] Khatib F, DiMaio F, Cooper S, Kazmierczyk M, Gilski M, Krzywda S, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nature structural & molecular biology. 2011;18(10):1175–1177.
[62] Nuwer R. Software could make rare diseases easier to spot. New Scientist. 2013;218(2913):21.
[63] White LW, Kowalewski TM, Hannaford B, Lendvay TS. SurgTrak: Evolution of a Multi-Stream Surgical Performance Data Capture System for the da Vinci Surgical Robot. Engineering and Urology Society. 2012;1. Ranked 5th out of 86 submitted abstracts by reviewers.
[64] White LW, Kowalewski TM, Hannaford B, Lendvay TS. SurgTrak: Synchronized Performance Data Capture for the da Vinci Surgical Robot. Hamlyn Symposium on Medical Robotics. 2012;1. Accepted for podium presentation.

[65] White LW, Kowalewski T, Hannaford B, Lendvay TS. SurgTrak: Affordable Motion Tracking and Video Capture for the Da Vinci Surgical Robot. In: Society of American Gastrointestinal and Endoscopic Surgeons, Proceedings of the 2011 Meeting of the SAGES, San Antonio, Texas. vol. 1; 2011. p. 204. Available from: http://www.sages.org/2011/resource/-posters.php?id=36030.
[66] Saha S. Appropriate degrees of freedom of force sensing in robot-assisted minimally invasive surgery. John Hopkins University; 2005.
[67] Lendvay TS, Brand TC, White L, Kowalewski T, Jonnadula S, Mercer LD, et al. Virtual Reality Robotic Surgery Warm-Up Improves Task Performance in a Dry Laboratory Environment: A Prospective Randomized Controlled Study. Journal of the American College of Surgeons. 2013;.
[68] on Quality of Health Care in America C. To Err is Human: Building a Safer Heath Care System. 2000;.
[69] Haynes AB, Weiser TG, Berry WR, Lipsitz SR, Breizat AHS, Dellinger EP, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. New England Journal of Medicine. 2009;360(5):491–499.
[70] Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General competencies and accreditation in graduate medical education. Health Affairs. 2002;21(5):103–111.
[71] Healy GB. The College should be instrumental in adapting simulators to education. Bulletin of the American College of Surgeons. 2002;87(11):10.
[72] Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. The American journal of surgery. 2010;199(1):115–120.
[73] Lendvay TS, Casale P, Sweet R, Peters C. Initial validation of a virtual-reality robotic simulator. Journal of Robotic Surgery. 2008;2(3):145–149.
[74] Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Annals of surgery. 2002;236(4):458.
[75] Peters J, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery. 2004;135(1):21–27.
[76] Gallagher AG, Ritter EM, Champion H, Higgins G, Fried MP, Moses G, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Annals of surgery. 2005;241(2):364.
[77] Mucksavage P, Lee J, Kerbl DC, Clayman RV, McDougall EM. Preoperative warming up exercises improve laparoscopic operative times in an experienced laparoscopic surgeon. Journal of Endourology. 2012;26(7):765–768.
[78] Anshel MH. The effect of arousal on warm-up decrement. Research quarterly for exercise and sport. 1985;56(1):1–9.
[79] Anshel MH, Wrisberg CA. Reducing warm-up decrement in the performance of the tennis serve. Journal of Sport and Exercise Psychology. 1993;15:290–290.
[80] Wrisberg CA, Salmoni AW, Schmidt RA. Warm-up effects in the learning of discrete motor skills. Acta Psychologica. 1975;39(4):311–320.

[81] Nascon J, Schmidt RA. The activity-set hypothesis for warm-decrement. Journal of Motor Behavior. 1971;3:1–16.
[82] Lee JY, Mucksavage P, Kerbl DC, Osann KE, Winfield HN, Kahol K, et al. Laparoscopic Warm-up Exercises Improve Performance of Senior-Level Trainees During Laparoscopic Renal Surgery. Journal of Endourology. 2012;26(5):545–550.
[83] Snijders TA. Power and sample size in multilevel linear models. Encyclopedia of statistics in behavioral science. 2005;.
[84] Tausch TJ, Kowalewski TM, White LW, McDonough PS, Brand TC, Lendvay TS. Content and Construct Validation of Robotic Surgery Curriculum Using an Electromagnetic Instrument Tracker. Journal of Urology. 2012;In Press.
[85] Schroeder D, Keefe D, Kowalewski T, White L, Carlis J, Santos E, et al. Visualizing Surgical Training Databases: Exploratory Visualization, Data Modeling, and Formative Feedback for Improving Skill Acquisition. 2012;.
[86] Hamilton CE, Mola WR, et al. Warm-up effect in human maze learning. Journal of experimental psychology. 1953;45(6):437.
[87] Anshel MH, Wrisberg CA. The Effect of Arousal and Focused Attention on Warm-Up Decrement. 1987;.
[88] Stefanidis D, Walters KC, Mostafavi A, Heniford BT. What is the ideal interval between training sessions during proficiency-based laparoscopic simulator training? The American Journal of Surgery. 2009;197(1):126–129.
[89] Arora S, Aggarwal R, Sirimanna P, Moran A, Grantcharov T, Kneebone R, et al. Mental practice enhances surgical technical skills: a randomized controlled study. Annals of surgery. 2011;253(2):265–270.
[90] Pugh CM. Warm-ups, Mental Rehearsals and Deliberate Practice: Adopting the Strategies of Elite Professionals. Journal of Surgical Research. 2012;176(2):404–405.
[91] Gallagher AG, Richie K, McClure N, McGuigan J. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. World journal of surgery. 2001;25(11):1478–1483.
[92] Gunther S, Rosen J, Hannaford B, Sinanan M, et al. The red DRAGON: a multi-modality system for simulation and training in minimally invasive surgery. Studies in health technology and informatics. 2007;125:149.
[93] Figert PL, Park AE, Witzke DB, Schwartz RW, et al. Transfer of training in acquiring laparoscopic skills. Journal of the American College of Surgeons. 2001;193(5):533.
[94] Zou G. A modified poisson regression approach to prospective studies with binary data. American journal of epidemiology. 2004;159(7):702–706.
[95] Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. 2006;.
[96] Jenison EL, Gil KM, Lendvay TS, Guy MS. Robotic Surgical Skills: Acquisition, Maintenance, and Degradation. JSLS: Journal of the Society of Laparoendoscopic Surgeons. 2012;16(2):218.
[97] Lin HC, Shafran I, Yuh D, Hager GD. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery. 2006;11(5):220–230.

[98] Makiyama K, Nagasaka M, Inuiya T, Takanami K, Ogata M, Kubota Y. Development of a patient-specific simulator for laparoscopic renal surgery. International Journal of Urology. 2012;19(9):829–835.
[99] Miller DC, Thorpe JA. SIMNET: The advent of simulator networking. Proceedings of the IEEE. 1995;83(8):1114–1123.
[100] Proctor MD, Bauer M, Lucario T. Helicopter flight training through serious aviation gaming. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology. 2007;4(3):277–294.
[101] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.
[102] Freund JE, Simon GA. Modern elementary statistics. vol. 12. Prentice-Hall Englewood Cliffs, New Jersey; 1967.
[103] Kowalewski TM. Real-time Quantitative Assessment of Surgical Skill. University of Washington; 2012.
[104] Chen C, White LW, Comstock B, Kowalewski T, Lendvay TS. Crowd-Sourced Assessment of Technical Skills (C-SATS): Faculty Experts vs. Amazon.com Mechanical Turk Project vs. Facebook. NEXTMED / Medicine Meets Virtual Reality. 2012;20.

Vita

Lee Woodruff White is the first and only son of Jane Woodruff Grant and Paul Howard White. They raised him and his sister Mariah in Eugene, Oregon where he earned an International Baccalaureate Diploma at the Eugene International High School and graduated from South Eugene High School in 2004. In 2008, Lee graduated Magna Cum Laude from the Tulane University of Louisiana with a Bachelor of Science in Engineering, earning Departmental Honors from the Biomedical Engineering Department. He joined the BioRobotics Laboratory where he was advised by Professor Blake Hannaford and Surgeon Thomas S. Lendvay. In 2013 he earned the Doctor of Philosophy degree in Bioengineering from the University of Washington. He will continue his education at Stanford University School of Medicine and hopes to pursue a career bridging engineering and surgery.