Received: 23 September 2019 | Revised: 5 March 2020 | Accepted: 15 March 2020

DOI: 10.1002/rcs.2105

ORIGINAL ARTICLE

The International Journal of Medical Robotics and Computer Assisted Surgery

WILEY

Systematic approach for content and construct validation: Case studies for arthroscopy and laparoscopy

Jake Farmer[1] | Doga Demirel[2] | Recep Erol[1] | Daniel Ahmadi[3] | Tansel Halic[1] | Sinan Kockara[1] | Sreekanth Arikatla[4] | Kevin Sexton[5] | Shahryar Ahmadi[6]

[1]Department of Computer Science, University of Central Arkansas, Conway, Arkansas

[2]Department of Computer Science, Florida Polytechnic University, Lakeland, Florida

[3]Pulaski Academy, Little Rock, Arkansas

[4]Kitware, Carrboro, North Carolina

[5]Department of Surgery, University of Arkansas for Medical Sciences, Little Rock, Arkansas

[6]Department of Orthopedic Surgery, University of Arkansas for Medical Sciences, Little Rock, Arkansas

Correspondence
Tansel Halic, Department of Computer Science, University of Central Arkansas, Conway, AR.
Email: tanselh@uca.edu

Funding information
National Cancer Institute, Grant/Award Number: 5R01CA19749; National Heart, Lung, and Blood Institute, Grant/Award Number: 1R01HL119248 - 01A1; National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Number: 1R44AR075481-01; National Institute of Biomedical Imaging and Bioengineering, Grant/Award Numbers: 1R01EB009362, 1R01EB014305, 1R01EB025241, 2R01EB005807, 5R01EB010037; National Institute of General Medical Sciences; National Institutes of Health, Grant/Award Number: P20 GM10342

Abstract

Background: In minimally invasive surgery, there are several challenges for training novice surgeons, such as limited field-of-view and unintuitive hand-eye coordination due to performing the operation according to video feedback. Virtual reality (VR) surgical simulators are a novel, risk-free, and cost-effective way to train and assess surgeons.

Methods: We developed VR-based simulations to accurately assess and quantify performance of two VR simulations: gentleness simulation for laparoscopy and rotator cuff repair for arthroscopy. We performed content and construct validity studies for the simulators. In our analysis, we systematically rank surgeons using data mining classification techniques.

Results: Using classification algorithms such as K-Nearest Neighbors, Support Vector Machines, and Logistic Regression we have achieved near 100% accuracy rate in identifying novices, and up to an 83% accuracy rate identifying experts. Sensitivity and specificity were up to 1.0 and 0.9, respectively.

Conclusion: Developed methodology to measure and differentiate the highly ranked surgeons and less-skilled surgeons.

KEYWORDS

arthroscopic rotator cuff, construct validation, content validation, gentleness, minimally invasive surgery, simulator, surgeon skill measurement, virtual reality

1 | BACKGROUND

Over the years, with extensive developments in fidelity of computer graphics and real-time interactivity, virtual reality (VR)-based simulators have become more widely used in medical education. On the contrary, conventional surgery education depending on human and animal models, cadavers, mannequins, and the apprenticeship model can be high-risk, non-repeatable, non-reusable, subjective, and very costly. VR simulation offers a safe and realistic visualization, with a clinically valid practicing environment that is reusable, and offers objective assessment metrics at a low cost. Due to these reasons, VR-based simulators became a critical tool for training surgeons in both teaching and training procedures including difficult surgeries where the field of view and hand motions are limited. VR-based simulators can help to train surgeons/residents for difficult minimally invasive surgeries such as arthroscopy and laparoscopy. For procedures such

as these, unnatural hand-eye coordination and a constrained field of view often can be confusing to new surgeons or residents.

Prior to widespread VR adoption, Dosis et al[1] attempted to objectively measure the hand movements and dexterity of surgeons using analog sensors and video synchronization. This showed that motion analysis for hand movements during surgery are a valid predictor of skill. Furthermore, VR-based simulators can be used for assessment of special skills that surgeons need[2] such as gentleness and respect for tissue handling. In order to quantify the smooth movement seen in surgery, as well as to discriminate between expert attending surgeons and novice surgeons, or between Post-Grad Year (PGY) 1-3, and PGY 4-5, movement features can be extracted from instrument motion data through the haptic(eg, touch) device and analyzed.[3] However, to differentiate the skill levels of surgeons, objective assessment is needed for validation of the VR simulators. Objective assessment is only possible with quantifiable data gathered from the VR-based simulators. The ultimate end goal of VR-based surgical simulators is to translate what is learned in the virtual environment to the real operating room.[4] This goal can be achieved through initial phase of preliminary validations such as content and construct validations, which the simulator must prove that the simulator content is a representative of the skills to be learned and also these skills could be discriminated for different skill levels of surgeons and physicians respectively.

In all surgeries, most specifically minimally invasive surgeries, gentleness and respect for tissue has precedence to avoid damaging tissue and prolonged recovery times.[5] Due to this ever-important factor, multiple surgical education programs have adopted some sort of simulation curriculum to fill in the gap between classroom learning and true apprenticeship and practice. As a part of this initiative, the American Board of Surgery (ABS) has even mandated that a box-trainer curriculum, the Fundamentals of Laparoscopic Training, be completed for certification. This curriculum must be completed prior to ABS certification as well.[6] Safe surgery is described as gentle handling of tissues, meticulous hemostasis, the avoidance of dead space, and adherence to impeccable surgical technique.[7] Particularly, gentle handling of tissues, or gentleness has become a valid part of operative performance assessment and ABS requires six clinical and operative performance assessments, specific to an area of specialty such as laparoscopic appendectomy, laparoscopic cholecystectomy.[8] Once certified, surgeons must be continuously trained to keep their certifications.[8,9] Gentleness itself has been assessed as a part of a procedural assessment by the ABS since 2012. Surgical skill has been shown to be a strong predictor of surgical outcome, and gentleness is one of the primary determinants of surgical skills.[5] Currently, surgical skill assessment is based on the apprenticeship model, which is mainly subjective and does not provide valuable feedback to the trainee. It has been shown that proficiency-based training involving a simulator is more effective, reducing operating room complications and errors.[10] Therefore, many general surgery programs have adopted a simulation-based curriculum, and it has been shown to improve performance.[11] However, this curriculum is a traditional approach of a "one size fits all" tactic, and does not address individual learning curves, nor does it allow for an objective approach to scoring. There are some commercially available simulators for laparoscopy and arthroscopy such as LapVR,[12] ArthroS,[13] and ArthroVR[14] to fill in the gap for the need for objective scoring, as well as allowing training in a consistent and reproducible fashion. However, one key area that they lack to assess is the gentle handling of tissues, or gentleness, which is our focus in this study. Gentleness is one of the principles described by William Halsted[15] as a component of "safe surgery" and has been assessed as part of a procedural assessment by the ABS[8] since 2012.

There are some attempts in literature to measure gentleness. Mackel et al[16] designed a pelvic exam physical simulator, called the E-Pelvis, to measure the forces while palpating the pelvic region. This included a physical mannequin of an adult female. The E-Pelvis samples data at 30 Hz from five pressure sensors. Therefore, it only measures applied pressure at certain points without providing additional feedback. Lamata et al[17] measured the force applied to pig tissue using force feedback, but only while pulling tissue. While pulling tissue is a common maneuver in arthroscopy and laparoscopy, it is not the only maneuver needed during these procedures.

In our attempt to measure gentleness and surgeon skill level including measuring for ambidexterity, we first compiled and identified significant feature set from simulator data and then used multiple classification algorithms are used to distinguish the two expert groups. The primary objective was to determine if the data provided from our simulators is effective to validate the content and construct of two virtual simulators, Gentleness Simulator and Virtual Arthroscopic Tear Diagnosis and Evaluation Platform (VATDEP),[18] and to distinguish the skill levels of surgeons as expert or novice. Skill level is identified using different clustering and classification algorithms such as K-means, Spectral Clustering, K-Nearest Neighbors, and Logistic Regression from the data received from the two simulators. Through our VR simulators and assessment, we are able to identify inexperienced surgeons who may need additional training earlier in the curriculum so that they can work on improving their skills.

2 | METHODS

In order to best measure gentleness of a surgeon, we have developed two unique 3D virtual reality scenarios: tennis racket and double grasper. These two scenarios and goals for each scenario were outlined and designed by the expert surgeons who are actively involved in the surgery residency training and periodically evaluate and assess residents' performance. In either of these scenarios, the user is presented with a simple task about gentle handling of tissue. While the user performs the task, the data from their hand movements is recorded via the Phantom Omni haptic devices. The simulators are developed using the Software Framework for Multimodal Interactive Simulations (SoFMIS) framework,[19] a multimodal, parallelized simulation framework that supports a high degree of customizability. The soft body dynamics are achieved through the use of Nvidia's PhysX SDK.[20] The simulator set up is shown in Figure 1.

We have also designed a shoulder arthroscopy diagnosis simulator, VATDEP[18] with the haptic feedback using Phantom Omni haptic

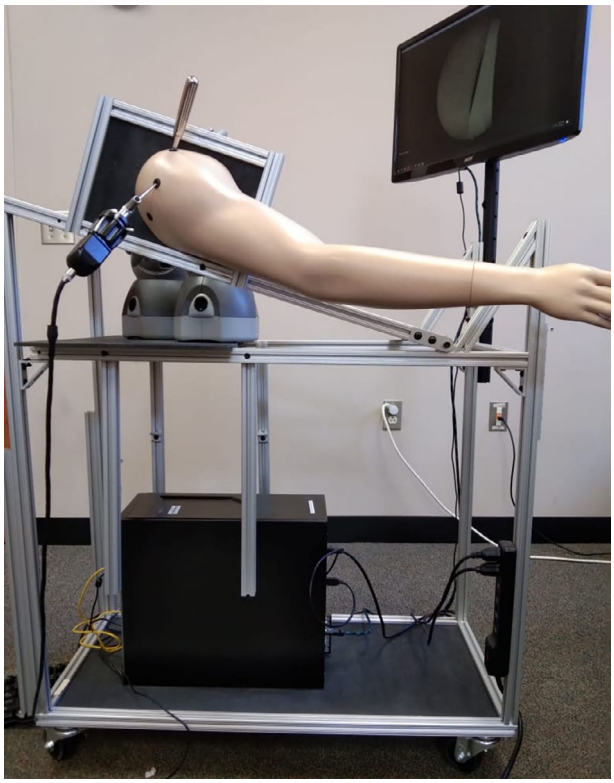

FIGURE 1 An overview of the simulator

devices. We presented two tasks to the user, which was to identify landmarks in the shoulder anatomy in a virtual scene, as well as to shave away simulated fur from the tear. The users were able to manipulate the camera and a probe or shaver tool in order to accomplish these tasks.

3 | GENTLENESS SIMULATOR

3.1 | Tennis racket task

For the tennis racket task, the user manipulates a tennis racket with the Phantom Omni haptic device. Main objective of this task is to measure the gentleness of the users contact with the balloon. In the scenario, any inadvertent and any severe handling could cause popping the balloon results in failing score or penalty in the case of permanent deformation. These threshold forces to pop the balloon are determined through expert surgeons' feedback through trials and errors. Balloon stiffness was set to 0.1 to correctly simulate the soft tissue behavior. Time that balloon stays in between two planes depends on the gentleness of the contact with the balloon. In the scene, there are three different colored planes. The floor plane is blue, and the green and yellow planes indicate the target region where the surgeon is expected to steadily keep the balloon for the maximum amount of time possible. The balloon is initially placed within the reach of the tennis racket, and the user is able to move the balloon upwards while getting force feedback from haptic devices in real-time. The user must keep the balloon above the green plane but below the yellow plane, or else both planes will flash red. In order to keep the balloon in the desired range, the surgeon must apply gentle forces, which are calculated based on the position of the racket and the weight of the balloon. The user also feels a force feedback to simulate a weighted balloon. This output force to the haptic is calculated based on the normal direction of the tennis racket when contact occurs, thus allowing a realistic feeling, including weight at an angle, across the entire output. This adds an additional challenge to this task, as the user must overcome the weight of the balloon, but not by so much that they fail the task. In addition to strong forces exerted, the balloon could also pop if it travels too far outside of the bounded region. Figure 2 shows the scene for the Tennis Racket task. The data that we recorded from the simulator included the position and velocity of the racket and the position of the midpoint of the soft body (balloon). In order to score well on this task, the trainee must keep the balloon in between the planes for as long as possible with as few hits as possible without any damage given to the balloon.

3.2 | Double Graspers manipulation task

The second task was the double graspers manipulation task. The main objective of the double graspers task is to use both hands to transfer the balloon, a soft body, from the right box side of the scene into the left box via the use of haptic devices. Balloon stiffness was set to 0.1 again to correctly simulate soft tissue like behavior. The user controls a pair of Phantom Omni haptic devices for this task and has the ability to open and close the jaws of the grasper at will with the buttons on the stylus of the haptic device. If the jaws of the grasper are closed too far on the balloon, or excessive force is applied to the balloon (vigorous shaking), then the balloon will pop, and the task must be restarted due to failing score. The user has an additional challenge imposed by having to move the balloon using the right-hand grasper from the right box towards the archway in the middle and grasp the balloon using the left-hand grasper and place it in the left box (or vice versa). If the user does not use adequate force to grasp the balloon, it can get away from the user and must be picked up again, which will cause an increase in task completion time. Figure 3 shows the scene for the double grasper task through the different stages: picking up, transferring, and placing.

4 | VIRTUAL ARTHROSCOPIC TEAR DIAGNOSIS AND EVALUATION PLATFORM

VATDEP is a simulator that we developed for training surgeons in the diagnosis of rotator cuff tears, as well as training them for minimally invasive surgery, specifically for treatment of rotator cuffs. The rotator cuff is a group of muscles and tendons located in the shoulder that connects the humerus (upper arm) to the scapula (shoulder blade). VATDEP is designed to allow the user to navigate around an anatomically correct model of the shoulder, including the correct ligaments,

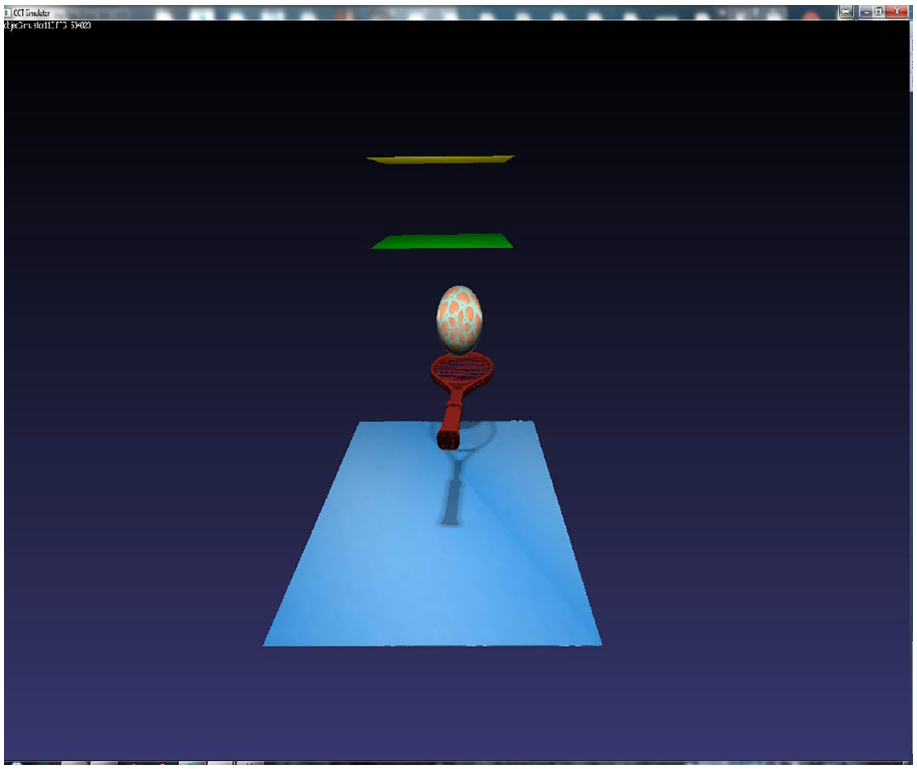

FIGURE 2 Tennis racket task, user tries to keep the balloon in between determined region without popping it by applying gentle forces

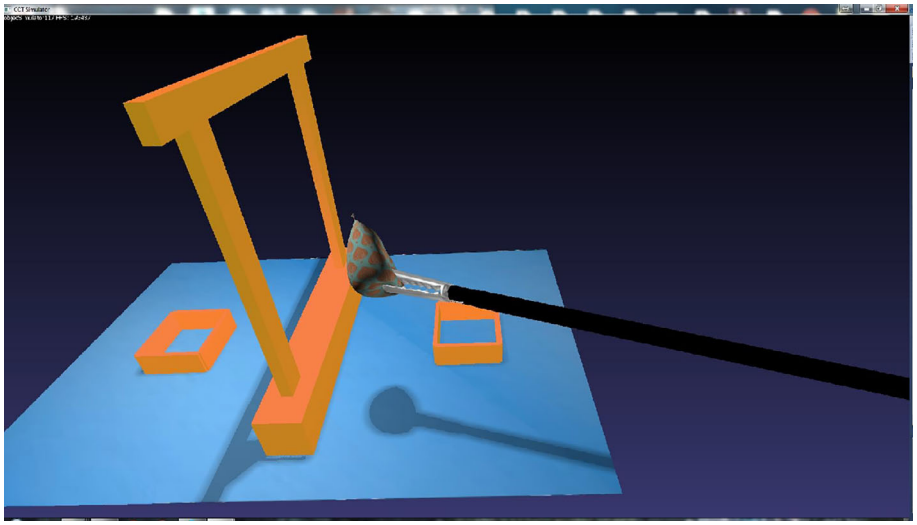

FIGURE 3 Double grasper task

tendons, and muscles. The user's task was to identify major landmarks in the scene, mark them, and continue on until all were found. In VATDEP, instructions were provided on screen for the users, as well as the name of the next landmark they must identify. The places where the pins were located are shown to the user, and the user had to apply smooth and gentle camera and tool movements to avoid getting lost in the scene or losing sight of their surgical tool. Both of these were common mistakes in arthroscopic surgeries and were noted by expert physicians as we were designing the simulator. Some aspects of the scene, such as the humeral head of the shoulder had force feedback, allowing the user to touch it and receive feedback through the haptic device. To achieve a high level of realism, physically-based rendering is used.[21] This allows for the integration of different lighting types and shading, such as bubbles during cleaning and heating, particle debris, depth of field, and many more special effects made possible with shaders. Figure 4 shows a sample VATDEP scene.

5 | DATA COLLECTION AND ANALYSIS

For both simulators, the study was conducted at the University of Arkansas for Medical Sciences (UAMS). For the Gentleness simulator,

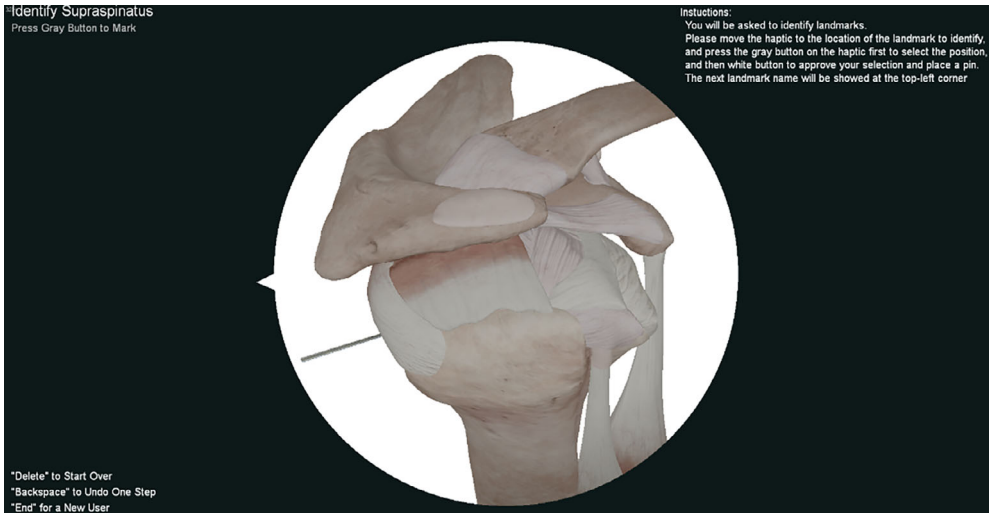

FIGURE 4 The Virtual Arthroscopic Tear Diagnosis and Evaluation Platform (VATDEP) scene

TABLE 1 Measurable performance for each simulator

Simulator name	Measurable performance
Gentleness (tennis)	Movement of the racket, time balloon spends between planes, amount of hits, amount of pops,
Gentleness (double grasper)	Movement of each grasper, time of the task, if the balloon has popped
VATDEP	Location of pins, camera movement, tool movement

Abbreviation: VATDEP, Virtual Arthroscopic Tear Diagnosis and Evaluation Platform.

TABLE 2 Features for gentleness simulator

Gentleness simulator	
Feature	***P*-value**
Popped	<<.01
Turning angle left (Mean)	.004
Soft body path length	.04
Average left path length	.05
Right path length	.05
Left turning angle (Median)	.05
Median jerk	.06
Average soft body path length	.07
Left path length	.08
Acceleration right (Median)	.1

the study was conducted with the Department of General Surgery, and for VATDEP, the study was conducted with the Department of Orthopaedic Surgery. Both studies were reviewed and approved by IRB at UAMS. Parameters used for performance measurements for both simulators can be seen in Table 1. The participants for both studies were given pre-questionnaires to fill out before using the simulator, with questions such as age, gender, hand dominance, level of training (Post Graduate Year), years in practice, number of procedures performed, number of procedures observed, as well as in the last 6 months, video game experience, experience using virtual simulators such as Fundamentals of Laparoscopic Surgery (FLS) or other virtual simulator training platforms. Each user was assigned a number and that was used in lieu of names.

Immediately after the use of simulators, we asked them to fill out a postquestionnaire. The postquestionnaire asked them to rate the difficulty of the scene, the degree of realism, and the quality/usefulness of force feedback from the haptic devices. We left some open-ended questions about concerns and comments, as well as possible additions to the simulator itself.

For the Gentleness Simulator, we had 23 subjects, including 4 expert and 19 novice surgeons, 12 of which were in PGY 1-3, and 7 of which were in PGY 4-5. For VATDEP, we had 10 users, split evenly between PGY 1-3, and PGY 4-5. For VATDEP, we used PGY 1-3 as the novice group and PGY 4-5 as the expert group.

Position and velocity data, as well as the angle of the grasper jaws in the double grasper task in the Gentleness Simulator were recorded for all users at 100 Hz. For VATDEP, the position of the camera and tool were recorded, as well as the forces that were applied to the humeral head. All features were compiled for each user across the tasks and we selected the most statistically significant feature sets from a *t*-test between novice and expert surgeons with a $P <= .1$ as seen in Tables 2 and 3. It has been shown previously that features such as velocity, acceleration, turning angle, etc. are valid predictors of a surgeon's level of skill.[22] This value ($P <= .1$) was chosen to prevent underfitting the data as we had few users for each simulator task. We did not find enough significant attributes from the Tennis task in order to cluster and classify with that data. Therefore, it is not included in the results.

We then used the feature sets in Tables 2 and 3 with numerous clustering algorithms. We use clustering to verify the distinction between the data before passing it to the classification algorithms.

We tested all the data with K-Means, Mean Shift, Affinity Propagation, Spectral Clustering, Agglomerative Clustering, and BIRCH. For the classification algorithms, we used K-Nearest Neighbors, Logistic Regression, and Support Vector Machine (SVM) with linear and radial basis function (RBF) kernels. All data was normalized with Z-Score, min-max, and max absolute value normalization.

6 | RESULTS

We compiled all feature data set in Tables 2 and 3, extracted the features with the lowest *P*-values, and normalized the data. The data was then clustered and classified 100 times, with different test / train splits for the classification data, thus performing cross-fold validation of the classifier performance. Table 2 shows the features with $P < =.1$ that were selected for Gentleness, and Table 3 shows the features with $P < =.1$ that were selected for VATDEP.

TABLE 3 Features for VATDEP

Virtual Arthroscopic Tear Diagnosis and Evaluation Platform (VATDEP)	
Feature	***P*-value**
Mean tool velocity	.01
SD jerk (Camera)	.02
SD acceleration (Camera)	.02
SD velocity (Camera)	.02
Mean jerk (Camera)	.03
Mean acceleration (Camera)	.03
Mean velocity (Camera)	.04
Time taken	.04
Tool path length	.08
Camera path length	.08

7 | CLUSTERING RESULTS

Once the features were selected and the algorithms had run, we used multiple metrics to quantify the results. For clustering results, we used the Silhouette Score, Adjusted Rand Index, Fowlkes—Mallows Index, the Jaccard Score, and the Mutual Information Index. The best results were found for a number of clusters n = 2. Figures 5 and 6 show the graphs for Gentleness Simulator for min-max and absolute value normalization methods and each algorithm that was run (including the unnormalized data). In this case for Gentleness Simulation, we were able to get 48% accuracy in recognizing novice and expert surgeons with Spectral Clustering and min-max normalized data. Normalizing the data offer large gains in improvement over some of the algorithms, especially K-Means and Spectral Clustering. However, since the range of data is smaller, the Silhouette Score dropped in almost all instances. As noted with Gentleness Simulation, there are variations in the results, but in this case Agglomerative Clustering performed the best with max absolute value normalization, achieving over an 80% success rate of classifying the users as either PGY 1-3 or PGY 4-5. Figure 7 shows a similar graph for the VATDEP simulator.

8 | CLASSIFICATION RESULTS

To quantify the clustering results, we looked at the precision, recall, F1 score, and the average accuracy of 4 different classification types, K-Nearest Neighbors (n = 2), n is selected as 2 due to different experience classes; novice and expert, Logistic Regression, and SVM with both linear and RBF kernels. We normalized the data in the same way as above and obtained massive improvements in the K-Nearest Neighbors algorithm. Figures 8 and 9 show the results for Z-Score and Min-Max classification on the Gentleness Simulator.

We were able to classify the practicing surgeons at best 74% of the time but were able to get up to 100% success rate with the novices using K-Nearest Neighbors after the data was normalized. All normalized data runs had an average accuracy of at least 80% or above correct classification for the Gentleness Simulator. We were able to

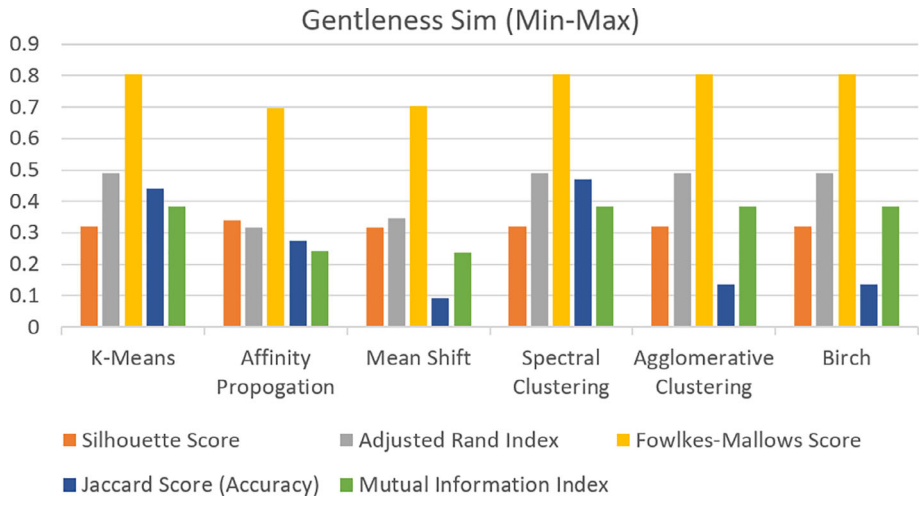

FIGURE 5 Gentleness Sim (Min-max normalized data)

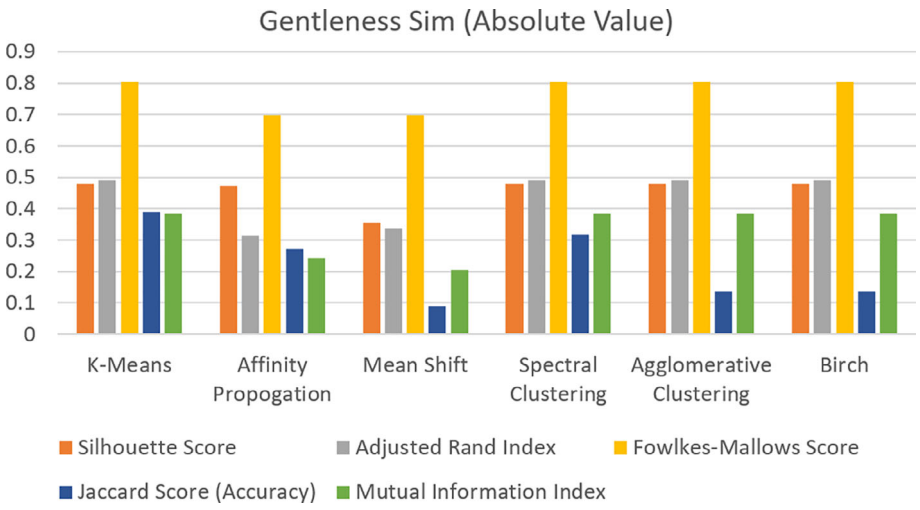

FIGURE 6 Gentleness Sim (Max absolute value normalized data)

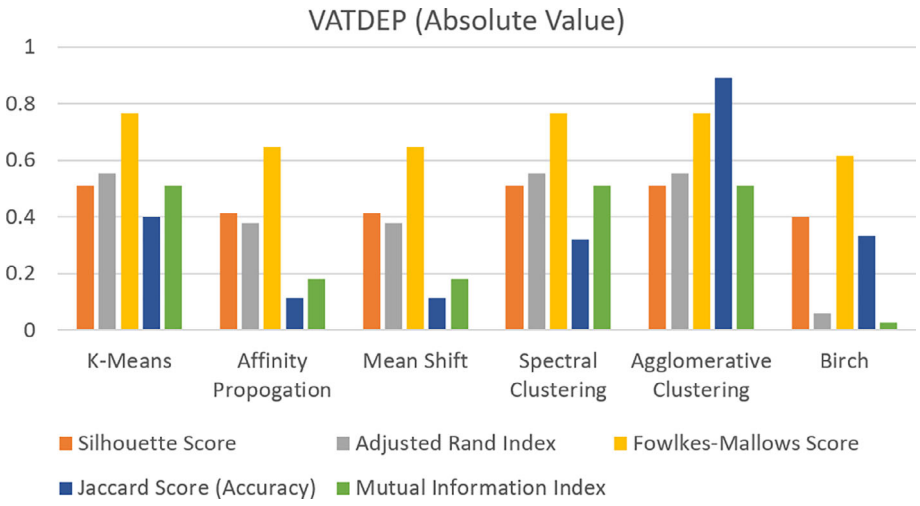

FIGURE 7 Virtual Arthroscopic Tear Diagnosis and Evaluation Platform (VATDEP) Sim (Max absolute value data)

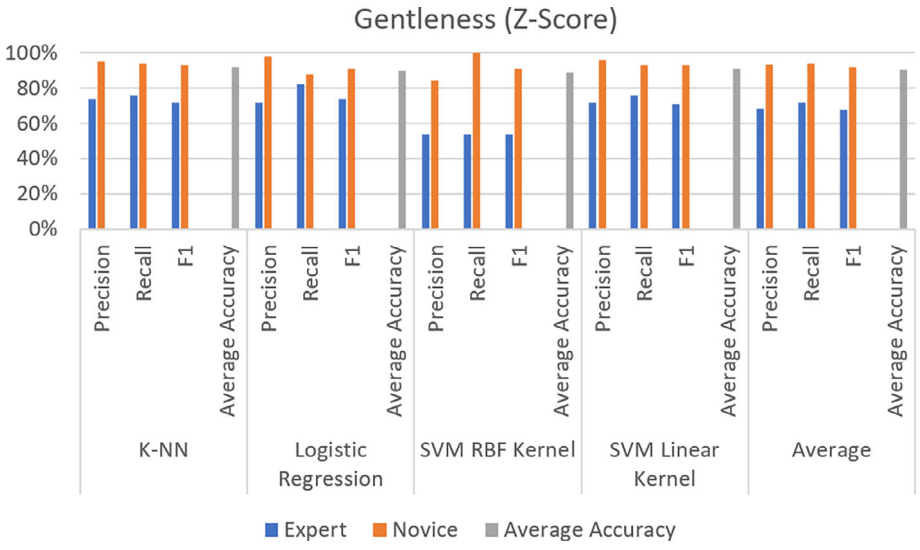

FIGURE 8 Gentleness classification (Z-Score normalized data)

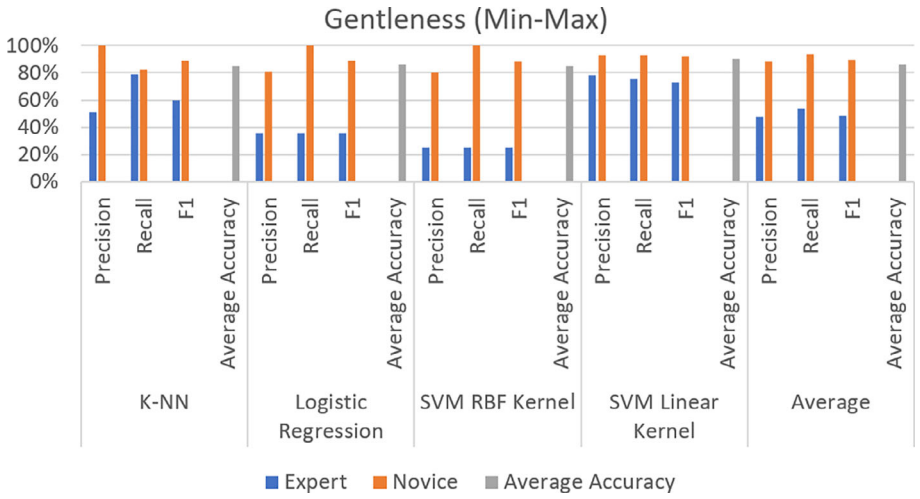

FIGURE 9 Gentleness classification (Min-max normalized data)

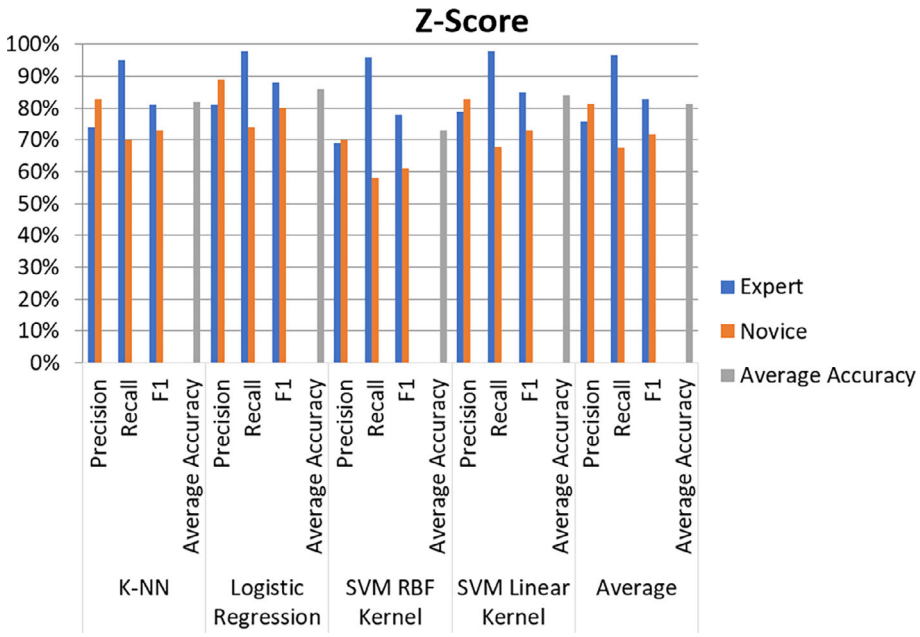

FIGURE 10 Virtual Arthroscopic Tear Diagnosis and Evaluation Platform (VATDEP) classification (Z-Score normalized data)

achieve up to 81% correct classification on the PGY 4-5 for VATDEP, and up to 89% correct classification on the PGY 1-3. Figure 10 shows the results for Z-Score from the VATDEP simulator.

9 | CONTENT VALIDATION RESULTS

In the post-questionnaires, we asked questions pertaining to the content validation of Gentleness and VATDEP and asked the users to rate the simulators on a Likert scale. Tables 4 and 5 summarize those results. For Gentleness, the experts highly rated its overall usefulness in enhancing gentleness at a 4.5, whereas the entire survey population, including the experts highly rated it at a 4.2. As a measurement of performance, the experts rated Gentleness at 3.25, and the entire population rated it at 3.1. For usefulness in improving laparoscopic skills, the experts rated it at 3.75, and the entire population rated it at 3.6. Finally, the experts highly rated the Gentleness Simulator at 4.5

TABLE 4 Gentleness simulator questions

Question	Expert mean score
Difficulty of double grasper task	4.5
Difficulty of tennis task	3.25
Object realism	3.25
Instrument realism	3
Overall realism	3
Quality of force feedback	3.25
Usefulness of force feedback	2.75
Usefulness in learning hand eye coordination	4.5
Usefulness in learning ambidexterity	4.5
Usefulness in improving laparoscopic skills	3.75
Usefulness as a measure of performance	3.25
Usefulness in enhancing gentleness	4.5

TABLE 5 VATDEP simulator questions

Question	PGY 4–5 Mean Score
Anatomical correctness of models	4.2
Rendering of models	4.4
Arthroscope manipulation	3.4
Probe manipulation	3.2
Force feedback	2.6
Size of tear	4.2
Location of tear	4.2
Type of tear	4.2
Overall realism	3.2

Abbreviations: PGY, Post-Grad Year; VATDEP, Virtual Arthroscopic Tear Diagnosis and Evaluation Platform.

for usefulness in learning ambidexterity, and the residents rated it at 4.1. For VATDEP, the PGY 4-5 users rated the size, location, and type of tear that was modeled at 4.2, and the survey population as a whole rated them at 3.8, 3.9, and 3.9, respectively. The PGY 4-5 users rated the overall realism of the simulator at 3.2, where the survey population rated it at 3.5.

10 | DISCUSSION

Our results demonstrate clear distinguishable groups amongst subjects for construct validation. However, since the sample size is small in both simulation validation study cases, our clustering results are not as expected. For example, the accuracy of K-Mean (.49) seemed to be quite low in the normalized data sets; however, upon inspection of the partitions produced from the K-Means clustering algorithm, we saw that only a few users were being misclustered, and those users were in the PGY 4-5 range, who are the closest to practicing surgeons in skill level.

In the double grasper manipulation task, user was allowed three attempts. Each attempt taking approximately 5 minutes, the first two attempts were for the user to get familiar with the task, haptic devices, and get accustomed to the 3D environment. Third attempt was used for actual performance where score is determined. On the other hand, tennis racket task was not standardized (attempt), but each user carried out the task for 5 minutes. For the VATDEP simulator, the users were given 5 minutes to get familiar with the haptic devices and the 3D environment. After the familiarization step, each user had one attempt.

Features such as the grasper angle and time taken to complete the task were not used in the case of the Gentleness Simulator because we did not find a statistically significant difference between the groups to include them in our final feature set; however, it exhibited significant measure for the VATDEP data set. Other insignificant features such as the grasper jaw angle for Gentleness or the identification of landmarks in VATDEP were also left out to eliminate noise. All features for the Gentleness Simulator were drawn from the Double Graspers Manipulation Task. The Tennis Racket task had very few significant features, such as the path length of the racket that could be used. This was due in part to the lack of consistent timing of the task during the study. Completion time for the task should have been held steady at 1 minute, but in some cases ran much longer.

In the VATDEP data set, a single user was being misclustered, and it was found that this user had prior experience with virtual simulators. We removed one user from the VATDEP data due to inconsistent and incomplete data. We also removed one user from the Gentleness Simulator data due to the lack of right-hand haptic data due to a glitch, as there was no data across any trials for this user.

While looking at features for selection from the Gentleness Simulator dataset, we noticed that almost all novice surgeons popped the balloon with their left hand during the double grasper task, while none of the expert surgeons did. Overall, we had a much easier time identifying novices or PGY 1-3 as their movements were much less refined and easily stood out against others.

For VATDEP, we saw that the hierarchical clustering algorithms, such as Agglomerative Clustering and BIRCH gave us much better results due to the level of noise present in our data. Since we were comparing what could be called novice and intermediate surgeons, there was less of a distinction between their feature sets, contrasted with the Gentleness Simulator that had strong distinctions between the groups of users. VATDEP also worked the best with Z-Score standardization, which allowed us to infer that the features that were used are very well distributed.

For the Gentleness Simulator, we saw that Spectral Clustering and K-Means gave very similar results, if not identical results, as they work in a very similar manner. Spectral Clustering first performs the eigenvalue decomposition, on the given data set, then performs K-Means Clustering on this reduced dataset. This indicated that the data is well separated, and a clear distinction is drawn between the two groups, as K-Means is very sensitive to noise and outlying data points. Min-Max and Maximum Absolute Value normalization worked the best with the Gentleness data. This indicated that there is a slight skew to the distribution, which is evident due to imbalance in data; we have more novice surgeons than experts.

We experimented other clustering methods (Affinity Propagation and Mean Shift) and attained overall low accuracy. We attribute this to the small datasets that we had to work with. However, methods based on hierarchical clustering and K-Means generated better accuracy. Therefore, the Gentleness data set is better suited to the types of algorithms that find median-based cluster centers, as opposed to mode-seeking algorithms such as Mean Shift. Sensitivity and Specificity were also computed for the results. Sensitivity was in the .9 to 1.0 range, and specificity was in the .8 to .9 range overall clustering and classification algorithms.

For both VATDEP and Gentleness, all of the classification algorithms generated similar accuracies with Z-Score standardization on classifying novices and experts. In both data sets, we can see the weakness of the RBF kernel of the SVM. With such a small dataset we unfortunately over- or under-fitted the data, and the precision suffered, especially in the case of the experts for the Gentleness

Simulator. All expert scores for the Gentleness Simulator were low because of how few there were, and the fact that the intermediate surgeons (PGY 4-5) would often be classified as expert. However, this was partially remedied by the normalization methods as shown in the results section above.

The most outstanding conclusion that we found was the prominent differences in the movement features between the left and right hands of the expert and novice groups. Often, we saw larger differences in the tests on the left hand for all data (both Gentleness Simulator and VATDEP), signifying that the experts have more mastery of ambidexterity than the novice surgeons. This is crucial during minimally invasive surgery as the surgeons must be able to use both hands equally well. We also found that surgeons were less focused on the movements of their non-dominant hand through features such as turning angle and acceleration. We believe that this is due to the confidence gained over several years of performing similar surgeries.

In post-questionnaire, we also asked face validation questions related to the realism of the simulators. From the Gentleness Simulator questionnaire, we found that even for such a simple scene, it was highly rated in terms of realistic handling as well as usefulness for training. Both tasks were rated as difficult by the users overall, and thus would be a unique training tool as they all completed both successfully. For VATDEP, the realism of the simulator was rated well, as the individual components were all scored at least 2.6 or above on the scale.

11 | CONCLUSION

In this paper, we introduced two distinct simulators, one for training and measuring the gentleness of a surgeon, as well as one for training surgeons in arthroscopic rotator cuff repair. We performed human subject studies to establish mainly content and construct validations at UAMS. During the validation studies, as a part of the data collection process, we have recorded kinematics data such as instruments' motion and haptic feedback and task-related data such as errors, correct identification of landmarks, and interaction details with the virtual scene (eg, balloon popping). We recruited a total of 33 subjects, 23 of which were for the Gentleness Simulator, and 10 were for VATDEP, ranging in experience from first-year medical students to attending surgeons with several years of experience. We performed post-questionnaire to get feedback from the subjects. We demonstrated that simulators are determined to be useful for enhancing certain skills such as tissue handling and ambidexterity. In addition, we have shown that with data collected from our simulators, we are able to distinguish between expert and novice surgeons based on their skill levels, as well as a multitude of other factors. Using a wide array of clustering and classification algorithms, we were able to show a distinction between two groups from both simulators. We also assessed the overall content validity of both simulators, and those who provided feedback were pleased with the current state of both simulators.

In the future, we plan to further perform validation studies such as learning curve, transfer of learning, and skill retention tests to establish the efficacy of the simulators for training and assessment. Upon the completion of these validations, we hope to incorporate metrics and clustering/classification techniques undertaken during the validations into the simulators directly so that the trainees can obtain quantitative feedback on their performances without expert intervention, which is time-consuming and costly. Our ultimate goal is to ensure that trainees could improve and attain the mastery in fundamentals and advanced skills in general and surgery specific procedures using VR based minimally invasive surgeries.

ACKNOWLEDGEMENTS

This publication was made possible by the Arkansas INBRE program, supported by a grant from the National Institute of General Medical Sciences (NIGMS), P20 GM103429 from the National Institutes of Health (NIH). This project was also supported by NIH/NIAMS R44AR075481-01, NIH/NCI 5R01CA197491 and NIH/NHLBI 1R01HL119248-01A1, NIH/NIBIB 1R01EB025241, 2R01EB005807, 5R01EB010037, 1R01EB009362, and 1R01EB014305.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

ORCID

Tansel Halic https://orcid.org/0000-0002-2558-4001

REFERENCES

1. Dosis A, Aggarwal R, Bello F, et al. Synchronized video and motion analysis for the assessment of procedures in the operating theater. *Arch Surg*. 2005;140(3):293-299.
2. Loukas C, Nikiteas N, Kanakis M, Georgiou E. Deconstructing laparoscopic competence in a virtual reality simulation environment. *Surgery*. 2011;149(6):750-760.
3. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. *Surg Endosc Interv Tech*. 2002;16(12):1746-1752.
4. Munz Y, Almoudaris AM, Moorthy K, Dosis A, Liddle AD, Darzi AW. Curriculum-based solo virtual reality training for laparoscopic intracorporeal knot tying: objective assessment of the transfer of skill from virtual reality to reality. *Am J Surg*. 2007;193(6):774-783.
5. Birkmeyer JD, Finks JF, O'reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369(15):1434-1442.
6. Fundamentals of Laparoscopic Surgery. Fundamentals of Laparoscopic Surgery. https://www.flsprogram.org/. Accessed April 10, 2019.
7. Cameron JL. William Stewart Halsted. Our surgical heritage. *Ann Surg*. 1997;225(5):445-458.
8. American Board of Surgery. www.absurgery.org. http://www.absurgery.org/default.jsp?index. Accessed April 10, 2019.
9. Williams RG, Sanfey H, Dunnington GL. A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. *Ann Surg*. 2012;256(1):177-187.
10. Angelo RL, Ryu RK, Pedowitz RA, et al. A proficiency-based progression training curriculum coupled with a model simulator results in the acquisition of a superior arthroscopic Bankart skill set. *Arthrosc J Arthrosc Relat Surg*. 2015;31(10):1854-1871.
11. Zendejas B, Brydges R, Hamstra SJ, Cook DA. State of the evidence on simulation-based training for laparoscopic surgery: a systematic review. *Ann Surg*. 2013;257(4):586-593.

12. Iwata N, Fujiwara M, Kodera Y, et al. Construct validity of the LapVR virtual-reality surgical simulator. *Surg Endosc*. 2011;25(2):423-428.
13. Stunt JJ, Kerkhoffs GMMJ, van Dijk CN, Tuijthof GJM. Validation of the ArthroS virtual reality simulator for arthroscopic skills. *Knee Surg Sports Traumatol Arthrosc*. 2015;23(11):3436-3442. https://doi.org/10.1007/s00167-014-3101-7.
14. Bayona S, Fernández-Arroyo JM, Martín I, Bayona P. Assessment study of insightARTHRO VR® arthroscopy virtual training simulator: face, content, and construct validities. *J Robot Surg*. 2008;2(3):151-158. https://doi.org/10.1007/s11701-008-0101-y.
15. Osborne MP. William Stewart Halsted: his life and contributions to surgery. *Lancet Oncol*. 2007;8(3):256-265.
16. Mackel T, Rosen J, Pugh C. Data mining of the E-pelvis simulator database: a quest for a generalized algorithm for objectively assessing medical skill. *Stud Health Technol Inform*. 2006;119:355-360.
17. Lamata F, Antolin M, Rodriguez S, Oltra A. Study of laparoscopic forces perception for defining simulation fidelity. *Stud Health Technol Inform*. 2006;119:288.
18. Demirel D, Yu A, Cooper-Baer S, et al. A hierarchical task analysis of shoulder arthroscopy for a virtual arthroscopic tear diagnosis and evaluation platform (VATDEP). *Int J Med Robot*. 2016;13. http://onlinelibrary.wiley.com/doi/10.1002/rcs.1799/full. Accessed March 6, 2017.
19. Halic T, Venkata SA, Sankaranarayanan G, Lu Z, Ahn W, De S. A software framework for multimodal interactive simulations (SoFMIS). *MMVR*; 2011:213-217. Amsterdam The Netherlands: IOS Press. https://books.google.com/books?hl=en&lr=&id=2YZtot_CN_YC&oi=fnd&pg=PA213&dq=sofmis&ots=I5LRqfVDfe&sig=UPFhQG1kC8ppiWYAX4O84CywW9M. Accessed September 19, 2016.
20. Nvidia DZ. Physx sdk. http://developerNvidiaComphysx-Downloads 2011.
21. Pharr M, Jakob W, Humphreys G. *Physically Based Rendering: From Theory to Implementation*. USA: Morgan Kaufmann; 2016.
22. Chmarra MK, Klein S, de Winter JC, Jansen F-W, Dankelman J. Objective classification of residents based on their psychomotor laparoscopic skills. *Surg Endosc*. 2010;24(5):1031-1039.

How to cite this article: Farmer J, Demirel D, Erol R, et al. Systematic approach for content and construct validation: Case studies for arthroscopy and laparoscopy. *Int J Med Robotics Comput Assist Surg*. 2020;1–11. https://doi.org/10.1002/rcs.2105