Medical & Biological Engineering & Computing (2024) 62:1887–1897
https://doi.org/10.1007/s11517-024-03053-8

ORIGINAL ARTICLE

Face, content, and construct validity of a novel VR/AR surgical simulator of a minimally invasive spine operation

Sami Alkadri[1,3] · Rolando F. Del Maestro[3] · Mark Driscoll[1,2]

Received: 14 June 2023 / Accepted: 2 February 2024 / Published online: 26 February 2024

Abstract
Mixed-reality surgical simulators are seen more objective than conventional training. The simulators' utility in training must be established through validation studies. Establish face-, content-, and construct-validity of a novel mixed-reality surgical simulator developed by McGill University, CAE-Healthcare, and DePuy Synthes. This study, approved by a Research Ethics Board, examined a simulated L4-L5 oblique lateral lumbar interbody fusion (OLLIF) scenario. A 5-point Likert scale questionnaire was used. Chi-square test verified validity consensus. Construct validity investigated 276 surgical performance metrics across three groups, using ANOVA, Welch-ANOVA, or Kruskal–Wallis tests. A post-hoc Dunn's test with a Bonferroni correction was used for further analysis on significant metrics. Musculoskeletal Biomechanics Research Lab, McGill University, Montreal, Canada. DePuy Synthes, Johnson & Johnson Family of Companies, research lab. Thirty-four participants were recruited: spine surgeons, fellows, neurosurgical, and orthopedic residents. Only seven surgeons out of the 34 were recruited in a side-by-side cadaver trial, where participants completed an OLLIF surgery first on a cadaver and then immediately on the simulator. Participants were separated a priori into three groups: post-, senior-, and junior-residents. Post-residents rated validity, median > 3, for 13/20 face-validity and 9/25 content-validity statements. Seven face-validity and 12 content-validity statements were rated neutral. Chi-square test indicated agreeability between group responses. Construct validity found eight metrics with significant differences ($p < 0.05$) between the three groups. Validity was established. Most face-validity statements were positively rated, with few neutrally rated pertaining to the simulation's graphics. Although fewer content-validity statements were validated, most were rated neutral (only four were negatively rated). The findings underscored the importance of using realistic physics-based forces in surgical simulations. Construct validity demonstrated the simulator's capacity to differentiate surgical expertise.

Keywords VR/AR surgical simulation; Face · Content · And construct validity; Physics-based haptic feedback

1 Introduction

Virtual reality (VR) surgical simulators have been rapidly adopted as a more objective method of training and evaluating surgical technical skills, especially when compared to conventional training methods [1, 2]. VR training modules provide safe and controlled training platforms that allow residents to further develop their surgical skills [3]. Furthermore, the ability to generate automated scoring systems further supports the notion of integrating VR simulator systems in the training and the objective assessment of surgical residents in performing procedures. VR simulators collect enormous sets of data pertaining to the psychomotor interactions of the user during the completion of the simulated tasks. Such data are often transformed into performance metrics that play an important role in training and assessing

✉ Mark Driscoll
mark.driscoll@mcgill.ca

[1] Musculoskeletal Biomechanics Research Lab, Department of Mechanical Engineering, McGill University, Macdonald Engineering Building, 815 Sherbrooke St W, Montreal, QC H3A 2K7, Canada

[2] Orthopaedic Research Lab, Montreal General Hospital, 1650 Cedar Ave (LS1.409), Montreal, QC H3G 1A4, Canada

[3] Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery, Montreal Neurological Institute, McGill University, 2200 Leo Pariseau, Suite 2210, Montreal, QC H2X 4B3, Canada

surgical trainees. Recent developments have coupled the VR systems with haptic technology, which allowed trainees to develop their "feel" of the procedure before performing in vivo surgeries. This haptic technology allows real-time force-feedback which enhances the authenticity of the training programs [3]. In fact, our group strives to demonstrate the potential benefits of incorporating accurate physics-based haptic technology on learning outcomes through detailed quantification of surgical forces [4].

Despite the advancements of VR simulators in the surgical field, spinal surgeries lagged behind other disciplines [3]. In particular, a clear gap was present in VR simulators for spinal minimally invasive surgeries; until recently, spinal simulation training was still in its infancy with very little research in the past two decades to create a spinal surgical simulator [3]. Moreover, the high demand of spinal surgeries led to continuous improvements of both the surgical techniques and the skills of the surgeons. Numerous efforts were directed to establish novel minimally invasive spine surgical procedures that enhance patient safety and recovery [5]. Coupling the high demand for novel minimally invasive spine surgeries (MISS) with the range of difficulty associated with spine surgery has led to the development of novel spinal VR simulators with haptic feedback [6, 7]. These simulator platforms can deconstruct complex surgical procedures such as the Oblique Lateral Lumbar Interbody Fusion (OLLIF) into discreet steps allowing trainees to concentrate on specific technical skills in need of enhancement rather than those already acquired [7–9]. The OLLIF surgery requires learners to master a broad spectrum of surgical techniques, and each of these components can be assessed and trained utilizing virtual reality simulators [7, 10]. One such system is the VR/AR training platform developed by our group to train orthopedics and neurosurgeons on a novel minimally invasive OLLIF surgery.

The promising preliminary results exhibited by VR surgical training systems further encouraged its adaptation to surgical curriculums [11]. However, proper fundamental validation studies of the simulator systems are required. More specifically, the utility of such simulators in effectively training and assessing surgical trainees must be established through foundational subjective and objective validation steps, namely, face, content, and construct validity.

Face and content validity are established using a questionnaire. Face validity is the extent to which the developed simulation environment mimics the real surgery, whereas content validity is the extent to which the developed system is representative of the skills required to successfully complete the real surgery [12]. Construct validity refers to the ability of the simulator to distinguish between different levels of surgical expertise [1, 13, 14]. It is an objective validation step that relies on the enormous sets of data generated from the interactions of the user during the simulated task. Such data are often transformed into surgical performance metrics that play an important role in not only establishing construct validity but also in training and assessing surgical trainees. The use of statistical analyses is the gold standard for establishing construct validity [1, 13, 14]. Statistically significant differences in the scores among experts and trainees on the generated surgical performance metrics highlight the ability of the simulator to adequately differentiate between levels of surgical expertise.

While recent literature reflects a growing interest in more advanced forms of validation, such as concurrent and predictive validity, there is a discernible gap in studies demonstrating concrete foundational face, content, and construct validations [15–17]. Concurrent and predictive validity, evaluate how closely the outcomes of a newly developed simulator align with those of an established gold standard and assess whether skills acquired on the simulator yield better results in real surgical settings, respectively. The current research aims to address this gap by focusing on and establishing the foundational validation steps. These initial validations are crucial as they establish the basic authenticity and educational relevance of the simulator, which is a necessary precursor to more complex forms of validation like concurrent and predictive validity [18].

Hence, the scope of our work is deliberately concentrated on face, content, and construct validity of a novel OLLIF surgical approach that has not been explored previously. Therefore, the generated surgical metrics used as part of the construct validity step are considered unique and novel as they describe aspects of this new surgical approach. Furthermore, the study sheds light on the impact of using accurate physics-based force feedback on surgical simulation training, an aspect that to the best of the authors' knowledge is not previously studied. Lastly, the novelty explored in this study also includes a unique face and content validation approach by making use of a side-by-side cadaver study where participants directly complete the surgical scenario on a cadaver followed by completing the same surgical operation on the simulator.

2 Material and methods

2.1 The VR/AR simulator and the simulated scenario

This study utilized a novel VR/AR surgical training system developed by McGill University in affiliation with CAE Healthcare and DePuy Synthes part of Johnson & Johnson. The surgical simulator under consideration is a physics-based simulator of a minimally invasive spine single-level fusion. The geometry of the surgical scenes in the simulator is reconstructed from patient-specific data. The simulation runs on a high-performance gaming laptop (i7-8750H) with Windows 10 operating system. Similar to the surgical reality, the rendered images are displayed on two flat panel monitors to match the interface in the operating room: a built-in monitor and an external touch screen monitor. The monitor on the left in Figure 1 provides general surgical guides including a recorded animation displaying how to operate instruments during a step. The other monitor is an interactive touch screen displaying the laparoscopic views of the surgical area with which the surgeons interact. Haptic feedback is provided from a combination of a six-degrees of freedom ENTACT W3D device and a benchtop model that includes 3D printed vertebrae components, also exemplified in Figure 1. This is conveyed to the surgeons hand via analog surgical tools interchangeably hooked up the haptic system.

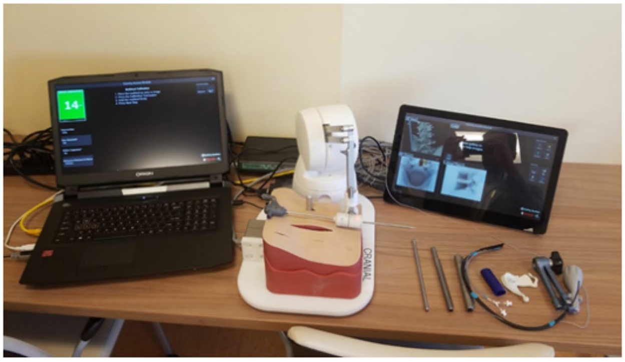


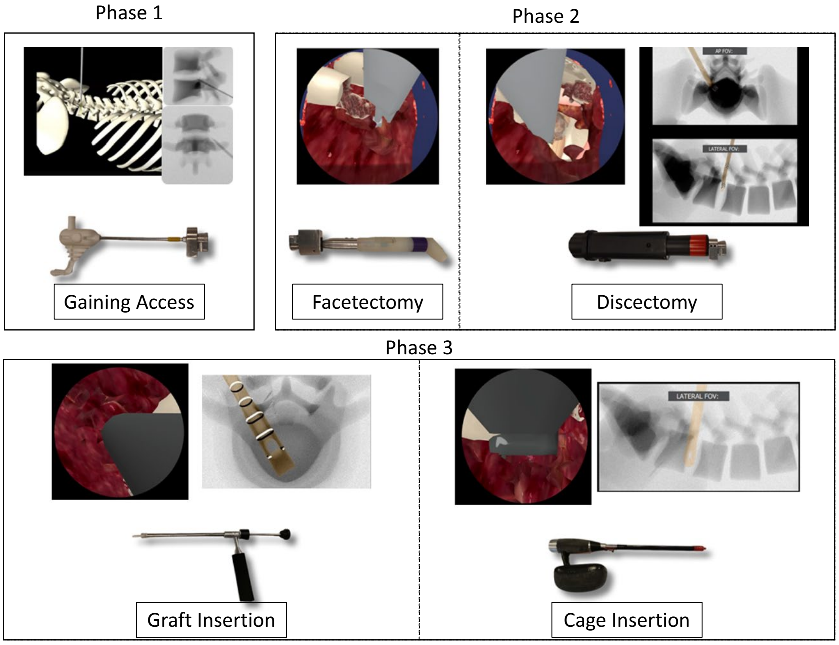
Fig. 1 The summarized simulator layout. Left is the laptop runs 120 Hz display, which indicates the instruction of the surgery process. The haptic device and benchtop model are in the middle. And right is the external display runs 60 Hz which indicates the four cameras that demonstrate the surgical area. The surgeon operates the haptic device based on the visual feedback from both monitors

The simulation focusses on three phases of the spinal surgery: gaining access through the back muscles, removing the intervertebral disk, and inserting graft and a spinal cage. The detailed steps along with the surgical tools used at each phase are demonstrated in Figure 2. The first phase of the simulated surgery includes the use of a multiprobe tool to gain access to the surgical area. Phase 2 requires the surgeon to first use a burr tool for drilling and performing a facetectomy, followed by using the Concord tool's suction mechanism to remove the remaining parts of the disk. Lastly, the surgeon is required to insert graft and a cage using the graft and cage insertion tools. The virtual volumetric model contains artificial muscle layers and an intervertebral disk, each providing realistic force feedback through interaction with the haptic device. The force feedback replicates the resistance provided by the instruments when penetrating through the muscles during an actual surgery using tailored mechanical properties. Prior to the start of the simulation, participants were made aware of all steps and instruments needed to complete the procedure via verbal and written instructions. No time limit was imposed on completing the simulated scenario.

2.2 Participants

Thirty-four participants were recruited to perform the virtual reality OLLIF scenario. Seven expert orthopedic surgeons out of the thirty-four participants were recruited in a side-by-side cadaver trial, where participants completed a minimally invasive spinal fusion surgery on a cadaver

Fig. 2 The three phases of the simulated surgery: Phase 1 includes gaining access to the disk using a multitool, Phase 2 includes facetectomy using a burr tool followed by a discectomy using a Concord Tool, and Phase 3 includes inserting graft followed by inserting a cage using the respective tools



and then immediately repeated the identical procedure on the surgical trainer/simulator. The remaining participants completed the trial without performing a cadaver surgery. All 34 participants were included in the face and content validity analyses. Due to errors during the simulation runs, only 24 individuals were included in the construct validity analysis: 10 post-residents, 5 senior residents, and 9 junior residents. Tables 1 and 2 present the demographics and the difference in experiences and knowledge of the 34 participants, respectively. The participants were divided into three groups: a post-resident group (3 neurosurgeons, 12 spine surgeons, 2 spine fellows, and 2 neurosurgical fellows), a senior-resident group (4 PGY 4–6 neurosurgery and 3 PGY 4–5 orthopedics residents), and a junior-resident group (4 PGY 1–3 neurosurgery and 5 PGY 1–3 orthopedics residents). This study was approved by an appropriate Research Ethics Board. All participants signed an approved written consent form prior to completing the simulation of the virtual spine surgery which took on average 1 h to complete.

Table 1 Demographics of the post-resident, senior-resident, and junior-resident groups

	Junior residents	Senior residents	Post-residents
No. of individuals	9	7	19
Sex			
Male	8	6	18
Female	1	1	1
Surgical specialty	Neurosurgery	Orthopedic surgery	
Level of training			
PGY 1-3	4	5	
PGY 4-6	4	3	
Fellows	3	2	
Consultants	2	12	

2.3 Face and content validity

All participants completed a questionnaire pertaining to face and content validity of the developed simulator using a 5-point Likert scale, where 1 indicated “strongly disagree” and 5 indicated “strongly agree.” There is no consensus on an acceptable median for sufficient face and content validity in the literature. In the current study, sufficient validity is assumed to be achieved if a median > 3.0 on a 5-point Likert scale is obtained for the post-resident group. Usually, face and content validity rely solely on the evaluations of the training system by expert surgeons [1, 12]. However, this study utilized responses made by non-experts (junior and senior-resident groups) to rate the consensus among experts and trainees on certain aspects of the simulator pertaining to both face and content validity [1, 12]. A Chi-square test was utilized to establish statistical significance of validity consensus. Comparing the consensus between the experts and trainees may be used to analyze the change in perspective with surgical experience [1]. This also allows for detailed analyses of validity that pinpoint aspects of the simulator that are adequately developed, require further improvements, or require a complete change [1].

The questionnaire was designed to gather detailed feedback from expert surgeons on two primary aspects: visual realism (face validity) and skill realism (content validity), evaluated within both the VR and AR dimensions of the simulation. Surgical and industry experts were consulted to ensure the questions were pertinent, clear, and effectively targeted the intended aspects of validity. For face validity, the questionnaire differentiated between the VR and AR components of the simulator, assessing graphical appearances of virtual anatomical structures and tools in VR, and the overall realism of the surgical setup in AR, including fluoroscopy, neuro-monitoring, and navigation tools. Content validity was similarly bifurcated, with VR questions examining the movements and haptic feedback of virtual tools, and AR questions focusing on the maneuverability

Table 2 Differences in previous experience, knowledge, and comfort level of the groups

	Junior residents	Senior residents	Post-residents
No. of individuals in each group who:			
Have previous experience using a surgical simulator	2 (22%)	5 (71%)	17 (89%)
Assisted on a TLIF	7 (77%)	7 (100%)	17 (89%)
Performed a TLIF solo	0 (0%)	0 (0%)	14 (73%)
Medina self-rating on 5-point Likert scale:			
Textbook knowledge of a TLIF	3.0 (3.0–4.0)	3.0 (3.0–4.0)	3.5 (1.0–5.0)
Surgical knowledge of a TLIF	3.0 (2.0–4.0)	3.0 (3.0–4.0)	3.5 (1.0–5.0)
Comfort level performing a TLID with a consultant in the room	3.0 (1.0–4.0)	4.0 (2.0–5.0)	4.5 (2.0–5.0)
Comfort level performing a TLIF solo	1.0 (1.0–2.0)	2.0 (1.0–4.0)	3.0 (1.0–5.0)

and tactile feedback of the physical tools. Additionally, the questionnaires incorporated elements from the side-by-side cadaver comparison study, an innovative aspect of our research. In this study, 7 expert surgeons from DePuy Synthes performed a transforaminal lumbar interbody fusion (TLIF) on a cadaver, followed by a simulation procedure. The subgroup completed the entire experiment within 1 h to ensure that the participants contrasted their experience on the virtual procedure to that on the cadaveric surgery in a side-by-side comparison. This direct comparison enabled the questionnaire to prompt participants, especially those involved in the cadaver study, to draw on their surgical experience and make direct comparisons between the simulator and the cadaveric procedure, ensuring a grounded and immediate tactile feedback assessment.

2.4 Construct validity

Construct validity was assessed using a priori metrics established independently for each module. During a simulation procedure, the surgical simulator recorded a series of data relating to the participants' use of the surgical tools. The collected data included variables such as position, time, and angles of the simulated surgical tools, as well as applied forces, removed volumes, and surgical tool contacts of specific anatomical structures. In total, 73 variables were collected throughout a simulation run. Subsequently, the recorded data were extracted and processed to generate surgical performance metrics that were used as a set of criteria to assess the performance of the participants in the virtual procedure. For example, position and time were combined to generate velocity metrics, forces, and contact detection were used to determine the forces used when removing anatomical structures, and position and contact detection were used to determine the path length used while interacting with anatomical structures. A total of 276 metrics were initially generated based on expert opinion, publications that focused on spinal fusion surgical performance, and novel metrics derived from the data. Subsequently, all derived metrics data were normalized using z-score normalization to reduce impact of outliers. Metrics were divided into three categories: motion, safety, and efficiency.

For all the generated surgical performance metrics, normality was tested using the Shapiro–Wilk test. For normally distributed data, variance homogeneity was further tested using Levene's test. To statistically measure the differences between the surgical groups, one of three statistical tests was used depending on the normality and variance homogeneity of the data. The standard ANOVA test was used if the data distribution was normal with equal variances across the groups. Welch ANOVA was used if normality was met but with heterogeneous variances. Lastly, Kruskal–Wallis parametric test was used for non-normally distributed data. A post-hoc Dunn's test with a Bonferroni correction was utilized to investigate differences between groups on significant metrics.

3 Results

3.1 Participants

Table 2 highlights the main differences between the groups based on previous experience, knowledge, and comfort levels performing and/or assisting in a TLIF (most similar procedure to simulated OLLIF). The senior-resident group (PGY 4 and higher) assisted in more TLIF surgeries and have a higher level of comfort assisting a TILF solo than the junior-resident group (PGY 1–3). Both the senior- and the junior-resident groups have no experience and low comfort in performing a TILF solo. Despite being the highest group having performed and assisted in a TLIF, the post-resident group ratings demonstrated that some recruited surgeons were non-spinal specialty and do not have textbook or surgical expertise in the TLIF surgery (median 3.5; range 1.0–5.0). In fact, 11% of the post-resident group have not performed or assisted in a TLIF previously.

3.2 Face and content validity

The face and content validity questionnaire consisted of 45 statements, 20 statements assessed face validity, and 25 statements assessed content validity. For face validity, post-resident group rated 13 statements positively (median > 3) and seven statements neutrally (median = 3) with no negatively rated statements (median < 3). For content validity, post-resident group rated nine statements positively (median > 3), 12 statements neutrally (median = 3), and four statements negatively (median < 3). The four negatively rated statements were all pertaining to interactions of the users with the burr tool. The median responses for each of the face and content validity statements are shown in Tables 3 and 4, along with the corresponding p values for a Chi-square test to assess the agreement in the response between junior, senior, and post-resident participants. All p values were greater than 0.05, indicating no significant differences among group responses.

3.3 Construct validity

Construct validity results showed significant differences between the three groups for eight metrics (Table 5). Box plots and pairwise comparisons of significant metrics are presented in Figure 3. The significant metrics spanned all three metric categories of motion, efficiency, and safety. Furthermore, the metrics differentiated the performance of

Table 3 Face validity median responses of the post-resident group with the Chi-square *p* values assessing inter-group agreeability

Validity statements	Post-residents median responses	Chi-square *p* value
The Physical Multitool accurately resembles the real surgical tool	4	0.356
The Virtual Multitool accurately resembles the real surgical tool	4	0.638
I am able to accurately set up the benchtop model to resemble a real surgery through the use of the Fox Arm and port	4	0.279
The orientation and angulation of the port in the physical world match what is seen in the virtual world	4	0.567
The Physical Bur accurately resembles the real surgical tool	4	0.807
The Virtual Bur accurately resembles the real surgical tool	3	0.177
The Physical Tissue Retractor accurately resembles the real surgical tool	4	0.331
The Virtual Tissue Retractor accurately resembles the real surgical tool	3	0.547
The Physical Concorde Clear accurately resembles the real surgical tool	4	0.627
The Virtual Concorde Clear accurately resembles the real surgical tool	4	0.487
The Physical Graft Delivery Device accurately resembles the real surgical tool	4	0.341
The Virtual Graft Delivery Device accurately resembles the real surgical tool	4	0.637
The Physical Cage Insertion Device accurately resembles the real surgical tool	3.5	0.1
The Virtual Cage Insertion Device accurately resembles the real surgical tool	4	0.511
The animation representing the cage insertion is similar to a real surgery	3	0.802
The visual guides shown during the simulation are similar to the ones used during a real surgery	4	0.586
The simulator system setup—including the positioning of the screen, the haptic device, and the benchtop model—is similar to a real surgical setup	3	0.515
The visual graphics shown in the Port Cam view are similar to reality	3	0.324
The internal impressions of the tissue model shown in the Port Cam view are similar to reality	3	0.554
The external impressions of the tissue model shown in the Port Cam view are similar to reality	3	0.67

the three groups while performing the most critical steps of the procedure.

4 Discussion

4.1 Overall validity

The newly developed VR/AR surgical simulator has been shown to attain face, content, and construct validity, making it a promising formative educational tool of a novel OLLIF surgical approach that has not been explored previously. Therefore, the generated surgical metrics used as part of the construct validity step are considered unique and novel as they describe aspects of this new surgical approach. The novelty explored in this study also includes a unique face and content validation approach by making use of a side-by-side cadaver study where participants directly complete the surgical scenario on a cadaver followed by completing the same surgical operation on the simulator. The study also gives an insight into the importance of using accurate physics-based force profiles in spinal surgical training.

4.2 Face and content validity

The results of the subjective validity assessment of the new surgical simulator show a high level of face validity with 13 out of the 20 statements reaching a median score greater than 3 on a 5-point Likert scale. The high number of positively rated statements and the lack of any negative feedback in the face validity questionnaire indicate that the system was perceived as having a good overall level of realism. Only seven statements were neutrally rated and did not reach validity (Table 3). Among the virtual tools displayed during the procedure, only the virtual bur tool and the virtual tissue retractor did not reach validity, with the rest of the tools in both the physical and virtual versions having sufficient validation in the face validity questionnaire. The rating of the appearance of the virtual bur tool might have been impacted by the negatively rated user experience of the tool. In fact, the only negatively rated statements in the content validity questionnaire were related to the interactions of the participants with the bur tool, which is further discussed in more detail later in this section. The virtual tissue retractor tool was the only tool with incomplete responses among participants; the use of the tissue retractor tool was optional during the simulation as in the case of the real surgery and

Table 4 Content validity median responses of the post-resident group with the Chi-square *p* values assessing inter-group agreeability

Validity statements	Post-residents median responses	*p* value Chi-square
I am able to maneuver the Multitool similar to a real surgery when puncturing on the model	4	0.527
The forces experienced using the Multitool during the gaining access step are similar to those experienced during a real surgery	3	0.689
The force difference between the soft tissue layers is appropriate	3	0.341
I can clearly distinguish between the soft and hard tissue	4	0.054
I am able to remove bone and soft tissue as needed to gain IVD access	4	0.536
I can clear an adequate access area	4	0.769
I am able to maneuver the Bur tool similar to a real surgery	2	0.051
The amount of bone removed using the Bur tool during each pass of the facetectomy step is similar to a real surgery	2.5	0.722
The bone forces experienced using the Bur Tool during the facetectomy step are similar to those experienced during a real surgery	2	0.158
The soft tissue forces experienced using the Bur Tool during the facetectomy step are similar to those experienced during a real surgery	2	0.42
I am able to use the Tissue Retractor Tool to protect the nerve similarly to a real surgery	3	0.546
The method of selecting annulotomy size is reasonable	3	0.541
I am able to remove the amount of soft tissue that I wanted	3	0.115
I am able to maneuver the Concorde Clear tool similar to comparable Curettes in a real surgery	4	0.527
The forces experienced using the Concorde Clear tool during the discectomy step are similar to those experienced using comparable Curettes during a real surgery	3	0.313
The torques experienced using the Concorde Clear tool during the discectomy step are similar to those experienced using comparable Curettes during a real surgery	3	0.319
I am able to remove IVD similar to a real surgery	3	0.494
I am able to scrape and prepare the endplates similar to a real surgery	3	0.274
I am able to tell how far into the IVD I have penetrated	3	0.421
The amount of disk removed as presented by the simulator metrics matches my expectations	3.5	0.302
When impacting on the Graft Delivery Device the changes at each mallet impact resemble a real surgical procedure	4	0.71
When impacting on the Cage Insertion Device, the changes at each mallet impact resemble a real surgical procedure	3	0.533
The movement of the Physical tools resembles a real surgical procedure as the graft is inserted in the IVD	4	0.456
The movement of the Physical tools resembles a real surgical procedure as the cage is inserted in the IVD	4	0.253
The overall tasks and the associated skills required to complete the simulation run are similar to those required to complete a real surgery	3	0.179

some participants chose not to utilize the tool, which may have contributed to the tool not reaching validity. Nevertheless, the physical versions of both the burr and the tissue retractor tools reached face validity, indicating that the graphics were not as effective in mimicking reality. In fact, six out of the seven neutrally rated statements were related to the graphics and animation, indicating that there may be room for improvement in this aspect of the simulation. However, refining the visual graphics and animations of a simulation negatively impacts the computational time per frame, which in turn impacts the ability of the simulator to maintain a realistic interactive experience [19]. When the frame rate per second becomes less than 20 Hz, discontinuous and lagging graphic feedback affects the user experience, which is related to the rate at which the brain processes visual data [19]. The current simulation was optimized to maintain the minimum required frame rate per second that ensures a realistic interactive experience while maximizing the realism of the graphics and animations [20]. The lack of any negative feedback in the face validity questionnaire supports the optimization decision and the fact that a good balance was found between realistic graphics and a realistic interactive experience.

Despite the relatively lower number of validated statements in the content validity questionnaire, the majority of the statements that did not reach validity were rated neutral and only four statements were negatively rated, which were all pertaining to the bur tool (Table 4). A recurring comment during the course of the trial was made regarding the use of a shield with the burr tool as demonstrated in Figure 4. The reduced depth perception of the camera view in the simulation coupled with the shield resulted in difficulties while handling the tool, which is demonstrated by the low median rating of the statement assessing the maneuverability of the

Table 5 Construct validity results showed significant differences between the three groups for eight metrics

Surgical step	Significant metrics	3 group split (junior vs. senior vs. post)		
		Data distribution	Variance homogeneity	Test statistic
Gaining access	**Total multitool tip path length**	Normal	Homogenous variance	ANOVA: $p = 0.032$
Facetectomy, discectomy, and Annulotomy	**Number of sign changes of the acceleration of the burr tool in the *Z*-direction**	Normal	Homogenous variance	ANOVA: $p = 0.022$
	Average jerk of the Concorde tool in the *Y*-direction	Non-normal	-	Kruskal–Wallis: $p = 0.04$
	Volume removed of the L4 endplate above the annulus fibrosus	Normal	Homogenous variance	ANOVA: $p = 0.041$
	Volume removed of the L5 endplate under the annulus fibrosus	Normal	Homogenous variance	ANOVA: $p = 0.042$
	Maximum force applied on the IAP using the burr tool	Non-normal	-	Kruskal–Wallis: $p = 0.036$
	Average distance to the nerve while operating the Concorde tool	Normal	Homogenous variance	ANOVA: $p = 0.03$
	Average distance to the Cauda while operating the Concorde tool	Normal	Homogenous variance	ANOVA: $p = 0.032$

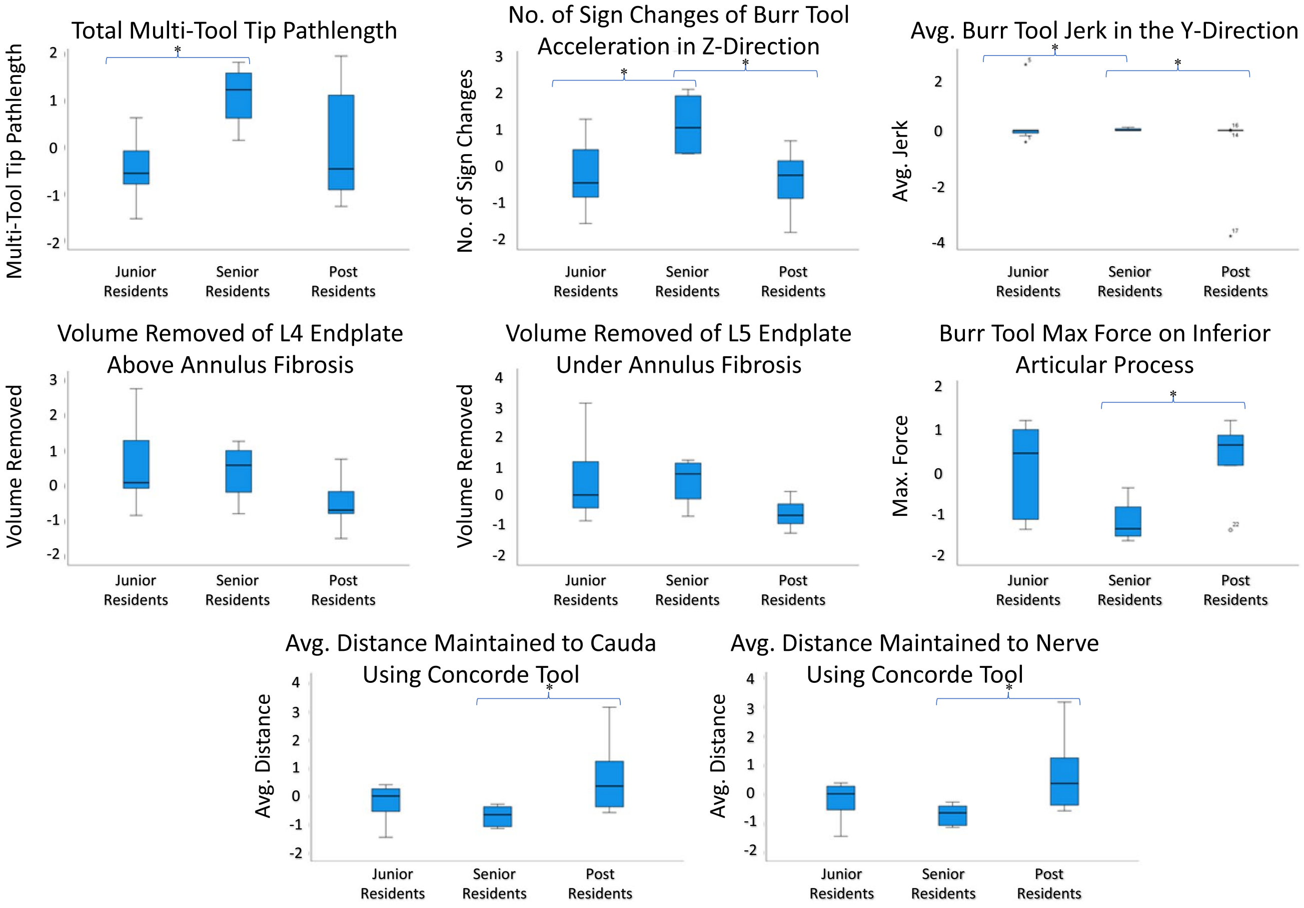


Fig. 3 Box plots and post-hoc Dunn's test with a Bonferroni correction of the 8 statistically significant metrics

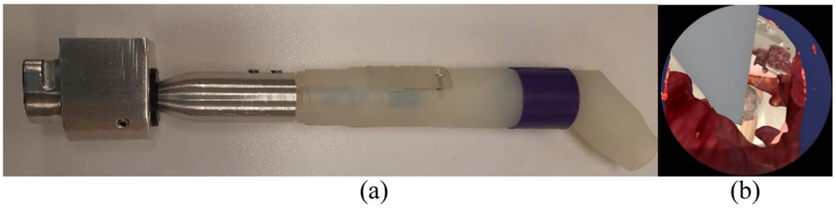


Fig. 4 **a** The physical bur tool. **b** Camera view of the virtual bur tool with a shield during the simulation

bur tool (Table 4). In general, one must be accurate in investigating the subjective rated aspects of a simulator system, as what can be perceived as a negative aspect of a simulator might be essential to capture reality. Careful investigation is required to determine if an added difficulty is representative of the skills required to complete the real surgery or if it is an unnecessary addition that needs refinement. While the overall graphics require further refinement, in the case of the current simulation, the reduced depth perception is essential to capture the true difficulties faced in the actual MI surgery. Therefore, the feedback obtained on the level of difficulty in handling the burr tool further supports the notion of face and content validity. Paradoxically, the importance of using realistic physics-based force profiles in surgical simulation is highlighted by the negatively rated statements regarding the forces experienced while operating the burr tool. The burr tool is the only tool in the simulation that is programmed with forces that are not based on cadaveric experiments. User interactions with the multitool and the Concorde tool that incorporated realistic forces based on cadaveric experiments were rated either positively or neutrally. This finding further supports the use of accurate physics-based force profiles in surgical simulations.

The Chi-square test was further used to assess the agreeability between group responses. For each statement, the null hypothesis was that the three groups had no differences in the ratings. All *p* values presented in Tables 3 and 4 had values greater than 0.05, failing to reject the null hypothesis and indicating that no statistically significant differences exist. These results support the notion that the groups were in agreement when assessing the aspects of the simulation.

4.3 Construct validity

The eight statistically significant metrics were derived from the surgical tools utilized in the most critical steps of the procedure (gaining access, facetectomy, and discectomy steps) and spanned all three metric categories. Developed through a collaboration of expert surgeons' insights and existing surgical literature, these metrics crucially demonstrate the simulator's ability to differentiate between various levels of surgical expertise, which is fundamental for construct validity. This differentiation suggests that the simulator can effectively measure the specific skills it intends to. Furthermore, these metrics have the potential to be teachable objectives for junior surgeons. They provide quantifiable targets in critical aspects of surgical performance, offering a pathway for skill development towards the benchmarks of more experienced surgeons. Thus, this construct validity analysis not only validates the simulator's assessment capabilities but also hints at its potential as a comprehensive training tool, which could significantly contribute to the advancement of surgical education.

During the gaining access step, the efficiency of the surgeons in reaching the surgical area represented by the multitool path length was significantly different among the groups. The results also indicated significant differences in handling the burr tool in the facetectomy step and the Concorde tool in the discectomy step and highlighted by the burr tool acceleration sign changes and the Concorde tool average jerk, respectively. Six safety metrics were identified during the facetectomy and discectomy steps. Metrics fall under the safety category if their effect results in either direct or indirect risk of injury or danger to the patient. Indirect safety metrics include unnecessary removals of anatomical structures such as the unnecessary removals of the L4 and L5 endplates identified in Table 5. Direct safety metrics include metrics that capture the preservation of important anatomy during the procedure, such as forces applied on anatomical structures and the proximity maintained to critical structures such as the nerve or the cauda. The maximum forces applied on the inferior articular process (IAP), and the average distances maintained to the nerve and cauda were identified as significant metrics in Table 5.

In general, a discontinuous learning pattern is characterized with non-sequential progression of skills while progressing from the junior-resident to the post-resident surgical level, passing through an inconsistent senior-resident level. Consider Figure 3, a clear discontinuous learning pattern can be seen in the motion and efficiency metrics. More specifically, both the post-residents and junior-residents were efficient with stable motions having seemingly smaller path lengths and less directional changes in their motion as compared to the senior-residents. Despite the similarity in the performances of the junior and post-residents in the motion and efficiency metrics, they are attributed to different reasons. The expert post-resident group utilizes precise and deliberate movements while the junior-residents are more reluctant and cautious. In the remaining metrics, it is not directly evident that a significant difference exists between the performances of the junior and senior-residents. The figure suggests that post-residents seem to remove less L4 and L5 endplates while being more wary of operating in critical proximity to the nerve and cauda when compared to the junior-residents and senior-residents.

The analysis done for construct validity is not just a validation of the simulator's effectiveness in distinguishing between different levels of expertise. It also lays the groundwork for its use as a comprehensive training tool, offering measurable and attainable goals for surgical skill improvement used for both training and assessment.

5 Conclusion

This study has established the face, content, and construct validity for the MI OLLIF simulated surgery on the newly developed VR/AR simulator. The unique side-by-side cadaver study and the use of accurate physics-based force profiles contributed in establishing the realism and educational value of the simulator. While some aspects, such as graphics and animation, could be improved, the system has been optimized to balance realistic graphics with a realistic interactive experience.

The face and content validity of the simulator were largely favorable, with only a few negative ratings. The majority of issues encountered were related to the virtual burr tool including the unrealistic force feedback for that particular tool as well as tool-handling difficulties. Upon further analysis, this feedback was shown not only to support the face and content validity of the simulator but also to highlight the importance of using realistic physics-based force profiles in surgical simulations as used for other surgical tools in the current simulation. The construct validity of the simulator is supported by the significant differences in performance metrics across different levels of surgical expertise. The analysis validates the simulator's ability to differentiate various expertise levels and establishes it as a comprehensive training tool, providing measurable goals for enhancing surgical skills in both training and assessment contexts. A discontinuous learning pattern was observed in the motion and efficiency metrics, with post-residents and junior-residents displaying seemingly smaller path lengths and fewer directional changes in their motion compared to senior-residents. In other metrics, post-residents demonstrated more precise and cautious behavior in terms of preserving important anatomy and maintaining safe distances from critical structures.

Overall, the VR/AR surgical simulator represents a promising formative educational tool for the OLLIF surgical approach. With further refinements and optimization, it has the potential to become an invaluable resource for training the next generation of surgeons in this innovative technique.

Funding NSERC Collaborative Research Development (CRD) Grant, Franco Di Giovanni Foundation, Brain Tumour Foundation of Canada Brain Tumour Research Grant, a Medical Education Research Grant from the Royal College of Physicians and Surgeons of Canada, and the Montreal Neurological Institute and Hospital.

Declarations

Ethical approval The institutional IRB approval was received for the study protocol and consent forms. This research involved recruiting surgeons to perform the virtual surgery on the simulator. Proper consent was obtained.

Competing interests The authors declare no competing interests.

References

1. Goldenberg M, Lee JY (2018) Surgical education, simulation, and simulators-updating the concept of validity. Curr Urol Rep 19(7):52. https://doi.org/10.1007/s11934-018-0799-7
2. Pfandler M, Lazarovici M Stefan P, Wucherer P, and Weigl M (2017) Virtual reality-based simulators for spine surgery: a systematic review. Spine J 17. https://doi.org/10.1016/j.spinee.2017.05.016
3. Vaughan N (2016) A review of virtual reality based training simulators for orthopaedic surgery. Med Eng Phys 38(2):59
4. El-Monajjed K and Driscoll M (2020) Analysis of surgical forces required to gain access using a probe for minimally invasive spine surgery via cadaveric-based experiments towards use in training simulators. IEEE Trans Biomed Eng 1. https://doi.org/10.1109/TBME.2020.2996980
5. McGill University to partner with industry in developing virtual-reality training platform for spinal surgery (2018). Available: https://www.mcgill.ca/newsroom/channels/news/mcgill-university-partner-industry-developing-virtual-reality-training-platform-spinal-surgery-287588
6. Alkadri S (2018) Kinematic study and layout design of a haptic device mounted on a spine bench model for surgical training. Mechanical Engineering, McGill University, Undergraduate Honours Program - Mechanical Engineering
7. Ledwos N, Mirchi N, Bissonnette V, Winkler-Schwartz A, Yilmaz R, Del Maestro RF (2020) Virtual reality anterior cervical discectomy and fusion simulation on the novel sim-ortho platform: validation studies. Oper Neurosurg 20(1):74–82
8. Cotter T, Mongrain T, Driscoll M (2022) Design synthesis of a robotic uniaxial torque device for orthopedic haptic simulation. J Med Devices 16(3):031008. https://doi.org/10.1115/1.4054344
9. Patel S, Ouellet J, Driscoll M (2023) Examining impact forces during posterior spinal fusion to implement in a novel physics-driven virtual reality surgical simulator. Med Biol Eng Comput 61:1837–1843. https://doi.org/10.1007/s11517-023-02819-w
10. Mirchi N et al (2019) Artificial neural networks to assess virtual reality anterior cervical discectomy performance. Operative Neurosurg 19(1):65–75. https://doi.org/10.1093/ons/opz359
11. Munz Y (2004) Laparoscopic virtual reality and box trainers: is one superior to the other? Surgical Endoscopy 18(3):485
12. Carter FJ et al (2005) Consensus guidelines for validation of virtual reality surgical simulators. Surg Endosc Other Interv Techn 19(12):1523–1532. https://doi.org/10.1007/s00464-005-0384-2
13. Huang C, Cheng H, Bureau Y, Ladak HM, Agrawal SK (2018) Automated metrics in a virtual-reality myringotomy simulator: development and construct validity. Otol Neurotol 39(7). https://doi.org/10.1097/mao.0000000000001867

14. Kwasnicki RM, Aggarwal R, Lewis TM, Purkayastha S, Darzi A, Paraskeva PA (2013) A comparison of skill acquisition and transfer in single incision and multi-port laparoscopic surgery. J Surg Educ 70(2):172–179. https://doi.org/10.1016/j.jsurg.2012.10.001
15. Stew B, Kao ST, Dharmawardana N, Ooi EH (2018) A systematic review of validated sinus surgery simulators. Clin Otolaryngol 43(3):812–822. https://doi.org/10.1111/coa.13052
16. Chawla S, Devi S, Calvachi P, Gormley WB, Rueda-Esteban R (2022) Evaluation of simulation models in neurosurgical training according to face, content, and construct validity: a systematic review. Acta Neurochir 164(4):947–966. https://doi.org/10.1007/s00701-021-05003-x
17. Van Nortwick SS, Lendvay TS, Jensen AR, Wright AS, Horvath KD, Kim S (2010) Methodologies for establishing validity in surgical simulation studies. Surgery 147(5):622–630. https://doi.org/10.1016/j.surg.2009.10.068
18. Søvik O et al (2023) Virtual reality simulation training in stroke thrombectomy centers with limited patient volume—simulator performance and patient outcome. Interv Neuroradiol. https://doi.org/10.1177/15910199231198275
19. Chen JY, Thropp JE (2007) Review of low frame rate effects on human performance. IEEE 37(6):1063–1076
20. Card SK (2018) The psychology of human-computer interaction. Crc Press

Sami Alkadri is a PhD student at McGill University who is exploring the validity of a surgical simulator aimed at improving surgical outcomes by way of identifying novel training metrics and objective ways to monitor and improve them.

Rolando F. Del Maestro is the William Feindel Professor Emeritus in Neuro-oncology at McGill University. As the Director of the Neurosurgical Simulation and Artificial Intelligence Learning Centre at the Montreal Neurological Institute and Hospital at McGill University, his research is focused on surgical simulation using the NeuroTouch platform. The research goal of his team is to improve global surgical education, surgical expertise, operative care, and patient outcomes utilizing simulation technology in cooperation with the National Research Council of Canada and multiple national and international research groups.

Mark A. Driscoll is a Professor in Mechanical Engineering at McGill University who is an expert in spine biomechanics. Prof. Driscoll focuses on gaining a better understanding of the mechanistic onset and progression of spinal disorders and leverages these findings to develop and improve diagnostic or monitoring methods and also focuses on evaluating current medical device treatment thereof.