Construct validity of the LapSim: Can the LapSim virtual reality simulator distinguish between novices and experts?

K. W. van Dongen, E. Tournoij, D. C. van der Zee, M. P. Schijven, I. A. M. J. Broeders

Department of Surgery, University Medical Centre Utrecht, Heidelberglaan 100, P.O. Box 85500, 3508 GA, Utrecht, the Netherlands

Abstract

Background: Virtual reality simulators may be invaluable in training and assessing future endoscopic surgeons. The purpose of this study was to investigate if the results of a training session reflect the actual skill of the trainee who is being assessed and thereby establish construct validity for the LapSim virtual reality simulator (Surgical Science Ltd., Gothenburg, Sweden).
Methods: Forty-eight subjects were assigned to one of three groups: 16 novices (0 endoscopic procedures), 16 surgical residents in training (> 10 but < 100 endoscopic procedures), and 16 experienced endoscopic surgeons (> 100 endoscopic procedures). Performance was measured by a relative scoring system that combines single parameters measured by the computer.
Results: The higher the level of endoscopic experience of a participant, the higher the score. Experienced surgeons and surgical residents in training showed statistically significant higher scores than novices for both overall score and efficiency, speed, and precision parameters.
Conclusions: Our results show that performance of the various tasks on the simulator corresponds to the respective level of endoscopic experience in our research population. This study demonstrates construct validity for the LapSim virtual reality simulator. It thus measures relevant skills and can be integrated in an endoscopic training and assessment program.

Key words: Endoscopic surgery — Virtual reality — Simulation training — Assessment

The technological revolution of endoscopic surgery has posed new challenges in surgical education. The skill set required for endoscopic surgery is different from the skill set required for traditional "open" surgery because

Correspondence to: I. A. M. J. Broeders

of the different operating environment. Endoscopic surgery requires three-dimensional orientation in a two-dimensional representation of the operating scene, as well as handling of endoscopic instruments [5, 8, 9]. Although endoscopic skills can be developed in the operating room successfully, it may not be the most appropriate or efficient environment to acquire such skills, given the steep learning curve that surgeons experience [1, 7, 11, 12]. Furthermore, financial and ethical issues and limited residential work hours impose a need to provide technical skill training in laboratory setting.

For the purpose of developing endoscopic skills, virtual reality (VR) simulators have been developed. A unique advantage of VR simulators is that they are both a training tool and an assessment device. During training objective measurements of performance are registered by the VR simulator and stored in its database. The database, in turn, provides the trainer or assessor with factual information on trainee performance status, without the need of being physically present.

Before simulator implementation in the surgical curriculum, systematic objective validation is required. The first step in objective validation is establishing "face validity." Face validity is the degree of resemblance between the concept instrument, the VR simulator, and the actual construct, psychomotor training, as perceived by a specific (target) population (surgeons and trainees) [2, 14]. Face validity is established by measuring the degree to which surgeons and trainees believe in the purpose and merits of the simulation environment. After having established face validity for the simulator, the simulator must be tested for its "construct validity," the degree to which the results of the "training session" as performed by the trainee on the simulator reflect the actual skill of the trainee who is being assessed [2, 14].

The notion of incorporating virtual reality training into the surgical curriculum has been suggested only recently and therefore validation testing for simulation concepts is a very recent development [3, 4, 6, 10, 13–17].

For the LapSim VR simulator (Surgical Science Ltd., Göteburg, Sweden), construct validity has been tested in three separate studies that used different methodology and yielded different results [3, 4, 16].

The purpose of this study therefore was to establish construct validity for the LapSim virtual reality simulator.

Materials and methods

Participants

There were 48 participants in this study. Each participant was assigned to one of three groups depending on their level of experience in endoscopic surgery. Group 1 consisted of 16 student interns lacking any form of endoscopic surgical experience. Group 2 consisted of 16 surgical residents in training who had performed more than 10 but less than 100 endoscopic procedures. Group 3 consisted of 16 experienced endoscopic surgeons who had performed more than 100 procedures. None of the participants has had any prior experience with the VR simulator.

Apparatus and tasks

The LapSim VR simulator uses the Virtual Laparoscopic Interface (VLI) hardware (Immersion Inc., San Jose, CA, USA), which includes a jig with two endoscopic handles. The VLI interfaces with a 2600-MHz hyperthreading processor Pentium IV computer running Windows XP and is equipped with 256 RAM, a GeForce graphics card, and a 18-in. TFT monitor.

The system features LapSim Basic Skills 2.5 software (Surgical Science Ltd., Göteburg, Sweden), from the LapSim Basic Skills package, consisting of eight tasks. The knot-tying task, in our opinion, does not represent the actual procedure. Therefore, the following seven tasks were selected and were the objects of study: camera navigation, instrument navigation, coordination, grasping, lifting and grasping, cutting, and clipping and cutting.

Tasks

A description of each of the selected tasks and the test by which the skill of the participants was assessed is defined below. In addition, the parameters measured and registered from each training session are described as indicative of the participant's skill in a particular task. The ability of a participant to successfully execute the selected tasks within a reasonable time frame while causing as little tissue damage as possible was measured as the total number of events causing damage (#) and maximum depth of damage (mm).

The *camera navigation* module's purpose is to train the user to navigate a scopic camera by finding and focusing on a number of balls that appear at random in a virtual environment. The size and number of balls and the time and pattern of appearance can be varied. In addition, the camera angle (30°), field-of-view, and zoom size can be adjusted. Parameters measured are time, misses, drift, trajectory and angular path of the camera, and tissue damage (total times and maximum depth).

The *instrument navigation* module's objective is to accustom the user to maneuvering and positioning endoscopic instruments. A number of balls appear in the virtual environment and have to be touched by two endoscopic instruments (one controlled with the right hand and one with the left hand). Number and size of the balls and time and pattern of appearance can be varied. Camera position can be rotated and put into motion. Assessed parameters are left and right instrument time, misses, pathlength and angular path, and tissue damage (total times and maximum depth).

The *coordination module* combines the instrument and camera navigation modules and consequently mimics the situation in diagnostic laparoscopy. One hand holds the camera, the other holds an instrument. Virtual balls appear randomly and have to be found by the camera, picked up with the instrument, and delivered in a target area. The difficulty can be varied according to the instrument and camera navigation modules.

The *grasping* module teaches the user to grasp, position, and navigate an object using a grasper. An appendix-shaped object has to be grasped, stretched until it releases, and positioned into a target area, while alternating the right and left instruments. Object number, size, timeout, and placement can be changed. The target size is variable as well. Camera options can be varied according to the instrument navigation module. Parameters measured are the same as those in the coordination module.

The *lifting and grasping module* aims at training bimanual handling. While lifting a box-shaped object, an underlying needle has to be grasped and moved to a target area. Camera, object, and target configurations can be varied as in the other modules. Parameters are the same as described for instrument navigation.

The *cutting module* focuses on grasping and handling an object with care and cutting it using ultrasonic scissors. After grasping and stretching a vessel, which will be torn off and hemorrhage if not handled with care, a colored area appears on the vessel. This has to be grasped and burned using a foot pedal. The excised segment then has to be moved to a target area. Number, size, and timeout of the segments and stretch sensitivity of the vessel can be adjusted. Rip and drop failure are two additional parameters measured as compared to the aforementioned modules.

Training

Three programs were designed with increasing level of difficulty: beginner, intermediate, and advanced. The easiest level was the manufacturer's default settings. The configuration of the adjustable options in the advanced level are challenging even for experienced endoscopic surgeons (>100 endoscopic procedures). Objects are smaller, have time restraints, and the camera view can be unstable or based on a 30° view. The adjustable options of the intermediate level were configured between the configuration of the beginner and that of the advanced level. After one familiarization run, which includes all of the selected tasks on all three levels, to get acquainted with the software, the actual formal training session was started. The participants started with the easiest task and ended with the most challenging task.

Assessment

There were 178 different parameters measured in total, as discussed in the *Material and methods* section. The participants were ranked by score for each of the 178 parameters. The scores on these different parameters were stored per participant. A ranking for all parameters was conducted by classifying the scores of individual trainees in the top 25% (first quartile), the mean 50% (second and third quartile), or the bottom 25% (fourth quartile) If the score of a participant ranked in the first quartile, he or she was awarded 2 points; if the participant score ranked in the second and third quartile, he or she was awarded 1 point. The participant did not receive any points for ending in the fourth quartile. Consequently, the maximum score any participant could achieve was 365 points (2×178).

The parameters were clustered into three categories (Table 1): speed, efficiency of instrument handling, and precision/accuracy.
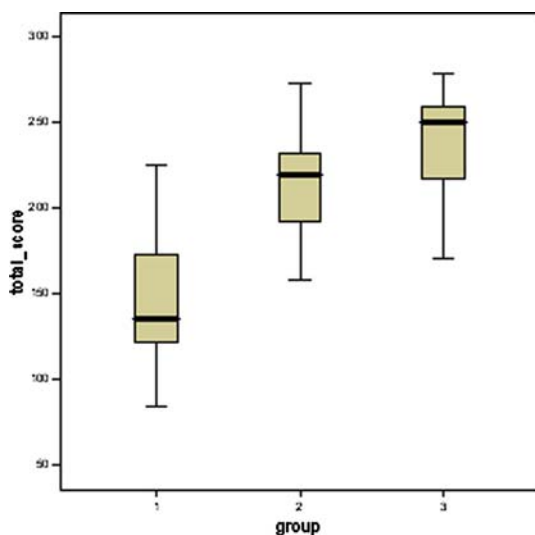
Table 1. Parameters per group

Speed	Efficiency	Precision
Time (s)	Path (m)	Tissue damage (n)
(instrument) misses (%)	Angular path (°)	Maximum damage (mm)
		Stretch damage (%)
		Incomplete areas (n)
		Bad clips (n)
		Dropped clips (n)
		Blood loss (L)

Fig. 1. Boxplot of total scores by the three groups.
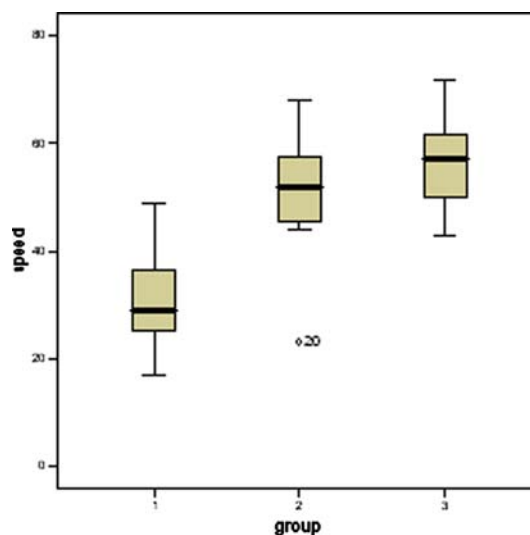

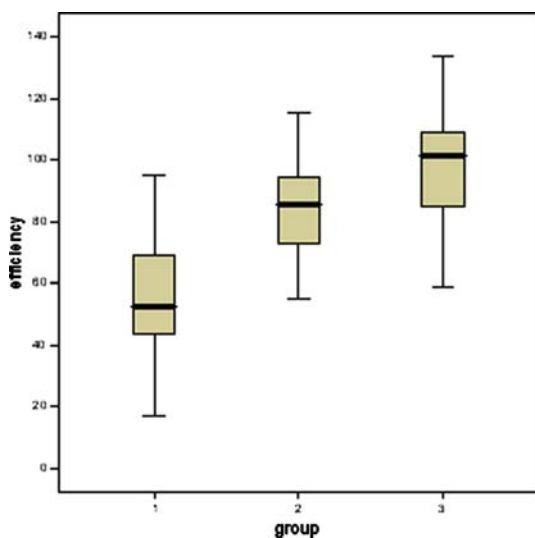**Fig. 3.** Boxplot of scores for speed by the three groups.

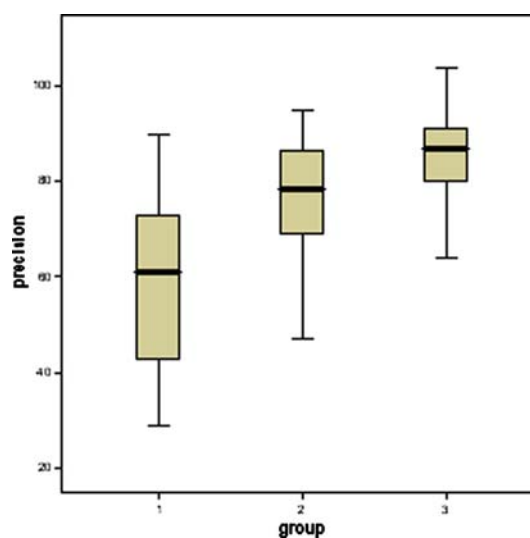**Fig. 2.** Boxplot of efficiency scores by the three groups.

Fig. 4. Boxplot of precision scores by the three groups.

Evaluation

All training tasks were evaluated for each level of difficulty (beginner, intermediate, advanced) and for their respective level of difficulty. Data analysis was done using SPSS v12.0 (SPSS, Inc., Chicago, IL). The one-way analysis of variance (ANOVA) with *post hoc* Tukey-Bonferroni test was used to determine differences in mean scores between the three groups where a $p \leq 0.05$ was considered statistically significant.

Results

We found that in general the higher the level of endoscopic experience of a participant, the higher the score. The differences between the groups are demonstrated at all three levels (beginner, intermediate, and advanced). At the advanced level the scores are most explicit and are therefore set out below.

Experienced surgeons (group 3) and surgical residents in training (group 2) showed statistically signifi-

cant higher scores ($p \leq 0.00$, $p \leq 0.00$) than novices (group 1) (Fig. 1), although the differences between the residents and the surgeons were not statistically significant ($p \leq 0.13$). Nevertheless, a trend in favor of group 3 was demonstrated.

The scores for efficiency, speed, and precision (Figs. 2, 3, and 4) are consistent with the overall score. Surgeons and residents demonstrate a higher score for parameters of efficiency ($p \leq 0.000$, $p \leq 0.000$), speed ($p \leq 0.000$, $p \leq 0.000$), and precision ($p \leq 0.000$, $p \leq 0.010$) than the inexperienced novices. The surgeons achieve higher scores than residents for all three parameters, although the differences are not statistically significant (efficiency, $p \leq 0.295$; speed, $p \leq 0.396$; precision; $p \leq 0.275$).

The standard deviation of all the scores is lowest in the group of surgeons, indicative of a smaller variability in outcome between participants in group 3 or a consistent experience level (Table 2).

Table 2. Means

Group	N	Speed		Efficiency		Precision		Total	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	16	30.4	9.04	55.31	20.72	60.75	18.55	147.81	40.99
2	16	51.56	10.37	85	17.15	76.63	13.98	214.63	30.09
3	16	56.63	8.51	96.5	19.77	85.44	9.4	239.56	29.53
Total	48	46.38	14.48	78.94	25.75	74.27	17.52	200.67	51.34

N = number of participants; SD = standard deviation

Discussion

This study demonstrates that the LapSim virtual reality simulator discriminates among participants of different endoscopic surgical experience, although the study has not tested the full range of skills and knowledge required to perform all varieties of endoscopic surgery. Specific objective end parameters (Table 2) that measure psychomotor skills were chosen as indicators for estimating actual endoscopic performance.

Establishing construct validity reflects the degree of empirical foundation of a concept instrument, e.g., the simulator [2, 14]. In practice, this is often established by measuring a logical difference in outcome between research populations with different levels of experience on a specific task of interest. Multiple studies have been conducted to validate different virtual reality systems as tools for training surgeons in endoscopic surgery skills [3, 4, 6, 10, 13–17]. These studies demonstrated construct validity for these systems. With regard to the relatively new LapSim virtual reality simulator, construct validity was investigated in three independent and separate studies [3, 4, 16].

Eriksen et al. [4] compared only two groups of surgeons: Group 1 (experienced) (>100 procedures, N = 10) and Group 2 (inexperienced) (<10 procedures, N = 14). Both groups performed all seven basic skills at an intermediate level, where the settings were configured to be challenging for an intermediate experienced endoscopic surgeon (>30 and <50 procedures). The parameters were analyzed separately. Time and efficiency parameters demonstrated statistically significant differences for all tasks. No statistically significant difference could be demonstrated for several of the error scores, in contrast with the present study. Residents and experts gained statistically significant higher scores for the combined error scores. The authors suggest that either small study size or poorly defined difficulty configurations were the cause of making these parameters nonvalid measures for surgical performance. These parameters could have been statistically significant if they had been combined into a similar relative scoring system as designed in the present study, or if they were linked to time for completion, as demonstrated by the "time-error" score of Sherman et al. [16]. Sherman et al. [16] demonstrated construct validity based on formulas that calculate a time-error score and a motion score. A total of 24 participants in three groups (7 naïve participants with no endoscopic surgical experience, 10 juniors with experience in <25 endoscopic procedures, and 7 experts with experience in >50 endoscopic procedures)

completed a training session of three tasks with increasing difficulty. The tasks were grasping, cutting, and clip applyication. The authors argue that time is not the exclusive indicator for a correct completion of a task. Consequently, they used time-error scores, which take both the time to complete a task and task-specific penalties into consideration. The results demonstrated statistically significant differences between the groups of participants for both scores. The task-specific scores, as constructed by Sherman et al. [16], are similar to our precision scores. In our study the standard deviation of the parameter "precision" shows the largest variability between the groups, e.g., novices to experts (18.5 vs. 9.5). Experts therefore appear to be more consistent in their performance than novices. Our results support the statement that accuracy is a concept that might not be addressed enough by the standard outcome parameters that are generated by the simulator. The parameter "speed" is both easy to measure and, in general, appealing to participants. Participants tend to prefer fast completion of a task over accuracy.

A time-error score appears to be an improvement in assessing performance compared with the standard manufacturers' end parameters.

The 54 participants in the study by Duffy et al. [3] executed basic skills tasks, with criteria based on manufacturer-recommended settings for individual exercises. There was no scoring system used, consequently the parameters were analyzed separately. Three groups of participants—junior residents (novices), senior residents (intermediates), and experts (surgeons)—were compared. Only a few parameters measured could discriminate between novices and experts.

The lack of a comprehensive scoring system, as designed for our study, limits the possibilities of demonstrating differences in performance between novices and residents. The most complex task (suturing) showed the most pronounced discrimination. A time-based analysis for task completion discriminated statistically significantly between novices and intermediates and between intermediates and experts. The authors conclude that their study demonstrated construct validity.

In our study the implementation of a scoring system enabled us to further assess the aspects of performance. Results demonstrate the importance of combining the different parameters. The assessment parameters of the simulator can be set according to individual preferences, thus providing opportunities to adjust for desired combinations or outcome parameters.

Coalescence of parameters seems useful as a reliable assessment of psychomotor skills. A combined scoring

system, set by experts, enables the creation of performance benchmarks that must be achieved by residents to achieve a predefined accreditation level.

Our results demonstrate that the registered performance scores show statistically significant differences between experts and residents vs. novices. Thus, in accordance with earlier studies [3, 4, 16], our study proves construct validity for the LapSim VR simulator. The LapSim psychomotor VR trainer can therefore be regarded as further established and empirically grounded.

To measure overall simulator performance based on these parameters, a relative scoring system was designed. This scoring system classifies a participant's performance on each of the measured parameters in percentiles and therefore relative to the overall research population. Because of the different measurement units of the parameters (seconds, millimeters, degrees), an overall scoring system is required to enable related parameters to be combined into one end score.

Limitations of the study

It must be stated that all three aforementioned studies, as well as our study, lack a power calculation for the group size. In retrospect, based on the results of time scores in the study of Duffy et al. [3], with a power of 0.8 and alpha set at 0.005, the group size should have been 17 instead of the chosen 16 persons per group.

Conclusion

This study demonstrated contruct validity for the LapSim virtual reality simulator. Our results showed that performance of the various tasks on the simulator indeed corresponds to the respective level of endoscopic experience in our research population. Provided that the other validation steps that need to be taken to complete the simulator's validation process are favorable, the LapSim VR simulator may be invaluable in training future endoscopic surgeons.

References

1. Bais JE, Bartelsman JF, Bonjer HJ, Cuesta MA, Go PM, Klinkenberg-Knol EC, van Lanschot JJ, Nadorp JH, Smout AJ, van der Graaf Y, Gooszen HG (2000) Laparoscopic or conventional Nissen fundoplication for gastro-oesophageal reflux disease: randomised clinical trial. The Netherlands Antireflux Surgery Study Group. Lancet 355: 170–174
2. Carter FJ, Schijven MP, Aggarwal R, Grancharov T, Francis NK, Hanna GB, Jakimowicz JJ: Working Group for Evaluation and Implementation of Simulators and Skills Training (2005) Consensus guidelines for validation of virtual reality surgical simulators. Surg Endosc 19: 1523–1532
3. Duffy AJ, Hogle NJ, McCarthy H, Lew JI, Egan A, Christos P, Fowler DL (2005) Construct validity for the LAPSIM laparoscopic surgical simulator. Surg Endosc 19: 401–405
4. Eriksen JR, Grantcharov T (2005) Objective assessment of laparoscopic skills using a virtual reality stimulator. Surg Endosc 19: 1216–1219
5. Gallagher AG, McClure N, McGuigan J, Ritchie K, Sheehy NP (1998) An ergonomic analysis of the fulcrum effect in the acquisition of endoscopic skills. Endoscopy 30: 617–620
6. Gallagher AG, Richie K, McClure N, McGuigan J (2001) Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. World J Surg 25: 1478–1483
7. Gouma DJ, Go PM (1994) Bile duct injury during laparoscopic and conventional cholecystectomy. J Am Coll Surg 178: 229–233
8. Hanna GB, Cuschieri A (1999) Influence of the optical axis-to-target view angle on endoscopic task performance. Surg Endosc 13: 371–375
9. Hanna GB, Shimi SM, Cuschieri A (1998) Randomised study of influence of two-dimensional versus three-dimensional imaging on performance of laparoscopic cholecystectomy. Lancet 351: 248–251
10. McCloy R, Stone R (2001) Science, medicine, and the future. Virtual reality in surgery. BMJ 323: 912–915
11. Moore MJ, Bennett CL (1995) The learning curve for laparoscopic cholecystectomy. The Southern Surgeons Club. Am J Surg 170: 55–59
12. Sariego J, Spitzer L, Matsumoto T (1993) The "learning curve" in the performance of laparoscopic cholecystectomy. Int Surg 78: 1–3
13. Satava RM (1993) Virtual reality surgical simulator. The first steps. Surg Endosc 7: 203–205
14. Schijven M, Jakimowicz J (2003) Construct validity: experts and novices performing on the Xitact LS500 laparoscopy simulator. Surg Endosc 17: 803–810
15. Schijven MP, Jakimowicz JJ, Broeders IA, Tseng LN (2005) The Eindhoven laparoscopic cholecystectomy training course—improving operating room performance using virtual reality training: results from the first E.A.E.S. accredited virtual reality trainings curriculum. Surg Endosc 19: 1220–1226
16. Sherman V, Feldman LS, Stanbridge D, Kazmi R, Fried GM (2005) Assessing the learning curve for the acquisition of laparoscopic skills on a virtual reality simulator. Surg Endosc 19: 678–682
17. Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A (1998) Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. Stud Health Technol Inform 50: 124–130