European Journal of Obstetrics & Gynecology and Reproductive Biology: X 4 (2019) 100092

Contents lists available at ScienceDirect

European Journal of Obstetrics & Gynecology and Reproductive Biology: X

journal homepage: www.elsevier.com/locate/eurox

Development and validation of a general and easy assessable scoring system for laparoscopic skills using a virtual reality simulator

JM Goderstad[a,*], L Sandvik[b], E Fosse[c,e], M Lieng[d,e]

[a] Department of Surgery, Sørlandet Hospital, Sykehusveien, 4838, Arendal, Norway
[b] Oslo Center for biostatistics and epidemiology, Oslo University Hospital, Norway
[c] The Intervention Centre, Oslo University Hospital, Oslo, Norway
[d] Department of Gynecology, Oslo University Hospital, Oslo, Norway
[e] Institute of Clinical Medicine, University of Oslo, Oslo, Norway

ARTICLE INFO

Article history:
Received 25 November 2018
Received in revised form 3 July 2019
Accepted 4 August 2019
Available online 13 August 2019

Keywords:
Laparoscopy
Simulation
Virtual reality
Procedural training
Assessment of surgical training

ABSTRACT

Objectives: To develop and validate a scoring system for laparoscopic skills for five specific tasks on a virtual reality simulator.
Study design: A longitudinal, experimental, non-randomised study including 30 gynecologists and gynecological trainees at three hospitals. The participants were categorized as inexperienced (Group 1), moderately experienced (Group 2), and experienced (Group 3).
The study participants performed ten repetitions of three basic skill tasks, a salpingectomy and a laparoscopic supracervical hysterectomy on a virtual reality simulator. Assessment of skills was based on time, error parameters and economy of movements measured by the simulator. We used the results (mean and SD for each parameter in all tasks) of the four last repetitions performed by the experienced gynecologists as the basic for the scoring system. Performance equal to, and higher than, this mean score gave 2 points. A decrease of 1 SD from the mean gave 1 point. Every score below gave 0 points. The mean score for the inexperienced, moderately experienced and experienced study participants was compared.
Results: The mean scores in Task 1 were 3.4 (SD 0.6) in Group 1, 3.4 (SD 0.6) in Group 2 and 5.1 (SD 1.1) in Group 3, respectively. There was a statistically significant difference in score between Group 1 and 3 ($p = 0.01$), and group 2 and 3 ($p = 0.01$). In Task 2 no statistical significant differences were found. In Task 3, the total mean scores were 1.7 (SD 0.7) in Group 1, 1.9 (SD 0.9) in Group 2 and 2.8 (SD 0.5) in Group 3, respectively. The difference in score between study groups was statistically significant when comparing Group 1 and Group 3 ($p < 0.01$) and Group 2 and Group 3 ($p = 0.02$).
In Task 4, the difference in used time between group 1 and 3 was statistically significant ($p = 0.03$). In task 5 there was a significant difference in performance score between group 1 and 3 ($p = 0.01$).
Conclusions: There was significant difference in scores between the experienced and the inexperienced gynecologist in four out of five tasks.
The scoring system is easy assessable and can be used for summative and formative feedback in proficiency-based assessment.

1. Introduction

Surgical competence is a combination of surgical technique, experience and strategy. As the surgical technology evolves, along with a heightened awareness of patient safety and concerns, surgical guidelines and surgical skill training programs become critical. Surgical training results in improved surgical performance when it comes to operating time, efficiency and safety [1–6].

Surgical competence can be measured and evaluated using a variety of different validated scoring systems, such as GOALS (Global Operative Assessment of Laparoscopic Skills) and OSATS (Objective Structured Assessment of Technical Skills), among others. Simulators and pelvic trainers are recommended and accepted training tools, and permit practice of surgical skills in a safe environment without compromising patient safety [7]. Access to pelvic trainers and simulators alone may not be sufficient to ensure an effective skills training [8]. A training program should be

* Corresponding author.
E-mail address: jeanne.mette.goderstad@sshf.no (J. Goderstad).

http://dx.doi.org/10.1016/j.eurox.2019.100092

based on proficiency and performance rather than a fixed number of repetitions or spent time [3]. Proficiency-based assessment is developed by experts performing the target procedure or a set of skills. The scores of the experts are then used to set the required score for passing the procedure or certification (summative feedback). Attaining basic psychomotor skills at an expert level on a surgical simulator system, could shorten the learning curves on real surgical procedures [1]. During development of surgical skills in the operating room, residents who have trained effectively on a surgical simulator are able to focus on the strategy of the procedure and the decision making during surgery, instead of focusing on the hand movements [9].

The objective of the study was to define and validate a scoring system for laparoscopic skills for five specific tasks on the simulator.

2. Materials and methods

This longitudinal, experimental, non-randomised study was conducted in accordance with the Declaration of Helsinki, and national and local regulations.

2.1. Inclusion

Gynecological trainees and consultants were invited to participate in the study. We recruited consecutively, until ten participants had been included in each study group. Prior to inclusion, all study participants received written information about the study, and they signed an informed consent for study participation. The surgical experience of each study participant at the time of inclusion was registered, and they were categorized into one of the following study groups:

Group 1: Inexperienced (performed less than 50 laparoscopic procedures, and previously never performed a laparoscopic hysterectomy)

Group 2: Intermediate experienced (previously performed more than 50 laparoscopic procedures, including more than five laparoscopic supracervical hysterectomies (LSH), but not performed total laparoscopic hysterectomies.

Group 3: Experienced (senior consultants performing total laparoscopic hysterectomy and surgery for women with deep infiltrating endometriosis).

2.2. Training

The training was carried out using the LAPmentor Express, Simbionix, 3D Systems, a portable, 2D non-haptic feedback simulator. At the first training session, all participants were given individual hands-on introduction to the simulator, and the tasks were presented. The program consisted of three basic skill tasks (Task 1, 2 and 3), a salpingectomy (Task 4) and a modified LSH (Task 5). All tasks were performed during each training session in a systematic order (Task 1–5, consecutively). This was repeated, dependent of available training time, up to maximum four times during one training session. The training was completed when all tasks had been performed ten times. The total training period was aimed to last between two and six weeks.

An instructor was present during all training sessions to assist the study participants in case they needed guidance on the simulator system or the tasks.

2.3. Description of the tasks

2.3.1. Task 1: two-handed maneuver

The task included exposure of nine balls embedded in jelly. A correctly exposed ball changed the color from red to green. All balls then had to be grabbed and placed into a basket.

This is a coordination task involving speed and precision. The objectives are to improve advanced bimanual skills, practice instrument manipulation and eye-hand coordination, and acquire tissue-handling skills.

The parameters measured were time (s), number of balls in the basket (n), total path length (cm) and instrument movement (number). In addition, number of errors was registered (only green balls should be grabbed).

2.3.2. Task 2: peg transfer

The participants lifted six objects from a pegboard with the left hand, transferred the object to the right hand, and placed them over the pegs on the pegboard. The process was then reversed.

The objectives are improved eye-hand coordination, use of both hands and depth perception.

The parameters measured were total time (s) and number of successfully moved objects (without loss and correctly placed on the pegboard) (n).

2.3.3. Task 3: pattern cutting

The participants used a grasper to apply traction exposing the best angle for the dominant hand to cut in the marked circle with accuracy.

The objective of this task is use of both hands and accuracy.

The parameters measured were total time (s) and errors (any deviation from the drawn line).

2.3.4. Task 4: left side salpingectomy

The participants used a grasper, scissors, and a bipolar forceps to remove the left tube. The total time used on the task (min) was registered. In case of an error (bleeding), it had to be corrected before commencing the salpingectomy.

2.3.5. Task 5: modified LSH

The participants were introduced to a step-by-step strategy starting on the left side and including [10].

1 Identification and division of the round ligament
2 Identification of the anterior leaf of the broad ligament and progressive cauterization of the ligament towards the middle medially paying attention to the bladder
3 Coagulation and division of the proper ovarian ligament and the fallopian tube
4 Division of the posterior leaf of the broad ligament
5 Identification, coagulation, and division of the uterine vessels
6 Step 1–5 was then performed at the right side
7 The cervix was exposed and the participant marked the correct level of amputation.

Total procedural time (min), total path length (cm), instrument movements (n) and errors (bleeding and improper respect of tissue/tissue handling) were registered. The registration started when the participant took hold of the left round ligament.

2.4. Statistics

All statistical tests were 2-sided, and $p \leq 0.05$ was considered statistically significant. Statistical analyses were performed using commercially available software (SPSS version 17.0; SPSS Inc., Chicago, IL). By using the findings in the publication by Fagerland et al, we found that the distribution of the variables for the parameters measured for each task (parameters of each task are described above) were sufficiently close to normal distribution for using the independent samples *t*-test [11].

2.5. Study sample size

The sample size calculation was based on the variable, "duration of task" from a procedural task (salpingectomy) in a study by Larsen CR et al. [9]. When calculating the sample size, we assumed that in the planned study, the difference in mean Total Time between the groups would be equal to the difference in median Total Time in the Larsen paper. The standard deviation of time in Larsens study is 90 s in the inexperienced group, and 40 s in the expert group. We assumed that similar standard deviation would be observed in our study. We furthermore assumed that the mean difference in time between Group 1 and Group 2 would be at least 90 s. It may be shown that if the true mean difference in time between these two study groups is at least 90 s, in a study with 80% test power and a significance level of 0.05, at least 10 physicians had to be included in each group. We consequently decided to include 10 study participants in each study group.

2.6. The scoring system

The results of the parameters of the four last repetitions of each task performed by the experienced participants were used as the base for the scoring system, and the mean and SD for these four repetitions was registered for each parameter in each task. In each repetition, a performance equal to, or higher than, the mean of the experienced participants in a registered parameter gave a score of two points. Up to one SD decrease from the mean in each parameter resulted in a score of one point. Every score below one SD gave 0 points. The scores from each parameter in each task (in Task 1, 2, 3, and 5) were added to give the total task score. Since the different tasks have different number of registered parameters, the maximum score differed between the different tasks (from 4 to 10), as illustrated in Figs. 1–3 and 5. In Task 4, we used time (min), as this was the only parameter in this task.

3. Results

3.1. Study participants

The mean age of the study participants in the three study groups was 31 years (SD 5.0) in Group 1, 36 years (SD 4.9) in Group 2 and 51 years (SD 7.3) in Group 3. All included study participants were right-handed. None of the study participants had any previous experience with the LapMentor simulator.

3.2. Training sessions

The study took place from September 2013 until May 2014. Some of the planned training sessions had to be postponed because of competing clinical activities and unexpected responsibilities. However, all study participants completed the training. The median total training period was 48 days (range 14–63 days) in Group 1, 19 days (range 7–61 days) in Group 2 and 25 days in Group 3 (range 4–60 days), respectively.

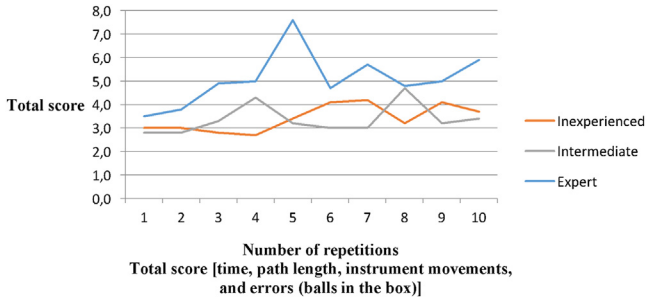

Fig. 1. Task 1 (Two-handed maneuver).

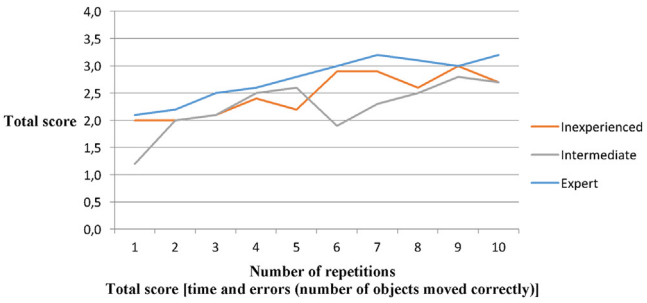

Fig. 2. Task 2 (Peg transfer).

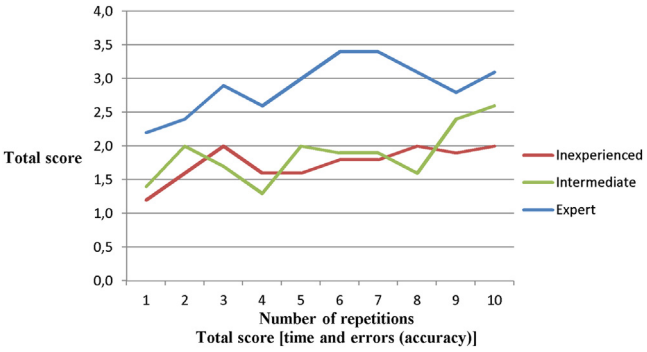

Fig. 3. Task 3 (Pattern cutting).

3.3. Scores of performance

The performance scores of all five tasks are presented in Figs. 1–5.

The mean scores in Task 1 were 3.4 (SD 0.6) in Group 1, 3.4 (SD 0.6) in Group 2 and 5.1 (SD 1.1) in Group 3, respectively. There was a statistically significant difference in score between Group 1 and 3 ($p = 0.01$), and group 2 and 3 ($p = 0.01$). The difference in score between Group 1 and Group 2 was not statistical significant ($p = 0.85$).

The total mean scores in Task 2 were 2.5 (SD 0.7) in Group 1, 2.3 (SD 0.7) in Group 2 and 2.8 (SD 0.3) in Group 3, respectively. In Task 2, no statistical significant differences in total score between the study groups were found (Group 1 vs. Group 3, $p = 0.1$, Group 1 vs. Group 2, $p = 0.5$ and Group 2 versus Group 3, $p = 0.1$).

The total mean scores in Task 3 were 1.7 (SD 0.7) in Group 1, 1.9 (SD 0.9) in Group 2 and 2.8 (SD 0.5) in Group 3, respectively. The

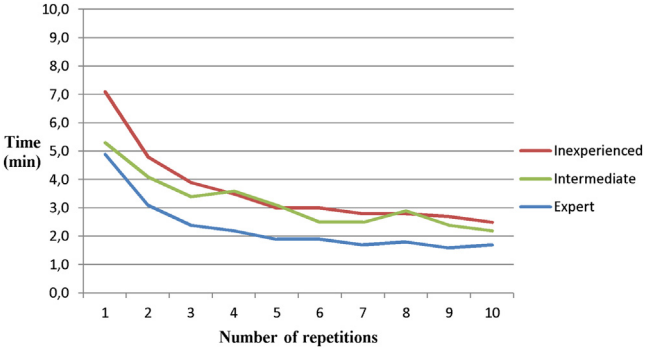

Fig. 4. Task 4 (Salpingectomy).

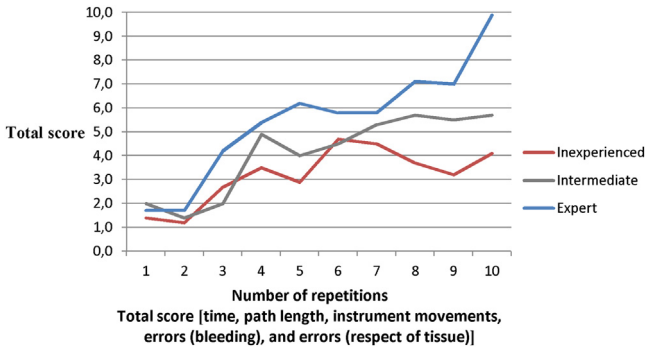

Fig. 5. Task 5 (Modified laparoscopic supracervical hysterectomy).

difference in score between study groups was statistically significant when comparing Group 1 and Group 3 ($p < 0.01$) and Group 2 and Group 3 ($p = 0.02$). There was no statistical significant difference in mean score when comparing Group 1 and Group 2 ($p = 0.60$).

The mean time used in Task 4 was 3.6 min (SD 1.4 min) in Group 1, 3.2 min (SD 0.9 min) in Group 2 and 2.3 min (SD 1.0 min) in Group 3, respectively. The difference in used time between group 1 and 3 was statistically significant ($p = 0.03$). There was no statistical significant difference in time when comparing Group 1 and Group 2 ($p = 0.45$) and Group 2 and Group 3 ($p = 0.06$).

The total mean performance score in Task 5 was 3.2 (SD 1.5) in Group 1, 4.0 (SD 1.6) in Group 2 and 5.3 (SD 1.8) in Group 3, respectively. There was a significant difference in performance score between group 1 and 3 ($p = 0.01$). The difference in mean score when comparing Group 1 and 2, and Group 2 and 3 was not statistically significant $p = 0.24$ and $p = 0.1$, respectively.

4. Comment

The results of this study showed a statistically significant difference in mean score when comparing the performance of experienced and inexperienced gynecologists in four out of five tasks in a standardized training program. Hence, the scoring system has validity for assessment of performance on the simulator.

The participants in Group 1 had some laparoscopic experience prior to inclusion. Inclusion of students in Group 1would probably resulted in larger differences between the groups. However, as the objective was to validate the scoring system for use in a clinical setting, we chose to include registrars that had started their laparoscopic training.

The participants in Group 2 were heterogeneous in respect to surgical experience. This might explain lack of significant differences between the groups in some tasks. The results of Group 2 are less relevant as the clinical importance of the scoring system is to differentiate between Group 1 and 3. Consequently, only comparing Group 1 and 3 probably would have improved the validation of the scoring system without reducing the quality of the study.

The lack of significant differences between groups in different tasks could furthermore be related to the level of difficulty of the tasks. The effect of training, measured as increase in total score, furthermore varied between the different tasks. This might be explained by the true value of the tasks. Given the results of group 1 and 2 in task 1, the value of this particular task can be questioned. The study participants were categorized into the study groups based on number and types of previous performed laparoscopic procedures. Previous authors have argued that previous performed procedures do not necessarily represent actual clinical competence [12–14]. Consequently, a different selection of study participants into the different groups might have influenced the mean performance and consequently the difference between the study groups. Learning curves express the relationship between an outcome variable, a score, and the number of repetitions of a given task, and can be used to determine when additional training is less likely to increase performance as well as to individualize training programs. Previous studies have investigated the influence on training schedules on surgical technical skills and show superiority for distributed training [15,16]. However, the interval between training sessions in a distributed training model may affect the outcome of the training. If the interval between the training sessions is too long, the retention of skills is influenced [16]. This may have been the case for some of the study participants especially in group 1. For practical reasons, ten repetitions of tasks were performed in our study setting. Brunner et al. have reviewed the literature describing the optimal number of repetitions during training [17]. Their data demonstrated that an initial plateau is reached after eight repetitions, but that the overall best score result was reached after 21–29 repetitions. Consequently, more repetitions in this study might have affected the training outcomes in all study groups. One of the strengths of virtual reality simulators is the standardized setting. This makes the setting fair for the participants, and can motivate them to participate in the study and complete the training period. This might also have contributed to all participants completing the training in this study. The results of this study demonstrate that different parameters in training tasks can be combined and integrated in a total score, which enables summative and formative feedback. The same principle can be done with GOALS, OSATS and other systems, but with these assessment tools an observer is mandatory to perform the assessment. Once the proficiency level of an exercise is set, it often will serve as a motivation factor for trainees wanting to increase their surgical skills.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank all the gynecologists who participated in the study, and their colleagues who covered their duties in the clinic while the training was performed.

References

[1] Larsen CR, Soerensen JL, Grantcharov TP, Dalsgaard T, Schouenborg L, Ottosen C, et al. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. BMJ 2009;338:b1802.
[2] Aggarwal R, Tully A, Grantcharov T, Larsen CR, Miskry T, Farthing A, et al. Virtual reality simulation training can improve technical skills during laparoscopic salpingectomy for ectopic pregnancy. BJOG Int J Obstet Gynaecol 2006;113(12):1382–7.
[3] Strandbygaard J, Bjerrum F, Maagaard M, Rifbjerg Larsen C, Ottesen B, et al. A structured four-step curriculum in basic laparoscopy: development and validation. Acta Obstet Gynecol Scand 2014;93(4):359–66.
[4] Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann surg 2002;236(4)458–63 Discussion 63-4.
[5] Kundhal PS, Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. Surg Endosc 2009;23(3):645–9.
[6] Gauger PG, Hauge LS, Andreatta PB, Hamstra SJ, Hillard ML, Arble EP, et al. Laparoscopic simulation training with proficiency targets improves practice and performance of novice surgeons. Am J Surg 2010;199(1):72–80.
[7] Aggarwal R, Ward J, Balasundaram I, Sains P, Athanasiou T, Darzi A. Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery. Ann Surg 2007;246(5):771–9.

[8] van Dongen KW, van der Wal WA, Rinkes IH, Schijven MP, et al. Virtual reality training for endoscopic surgery: voluntary or obligatory? Surg Endosc 2008;22(3):664–7.
[9] Larsen CR, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, et al. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. Surg Endosc 2006;20(9):1460–6.
[10] Goderstad JM, Sandvik L, Fosse E, Lieng M. Assessment of Surgical Competence: Development and Validation of Rating Scales Used for Laparoscopic Supracervical Hysterectomy. J Surg Educ 2016;73(4):600–8.
[11] Fagerland MW, Sandvik L. Performance of five two-sample location tests for skewed distributions with unequal variances. Contemp Clin Trials 2009;30(5):490–6.
[12] Bjerrum F, Strandbygaard J, Rosthoj S, Grantcharov T, Ottesen B, Sorensen JL. Evaluation of procedural simulation as a training and assessment tool in general surgery-simulating a laparoscopic appendectomy. J Surg Educ 2017;74(2):243–50.
[13] Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. Adv health Sci Educ: Theory Pract 2015;20(3):829–34.
[14] Korndorffer Jr. JR, Kasten SJ, Downing SM. A call for the utilization of consensus standards in the surgical education literature. Am J Surg 2010;199(1):99–104.
[15] Verdaasdonk EG, Stassen LP, van Wijk RP, Dankelman J. The influence of different training schedules on the learning of psychomotor skills for endoscopic surgery. Surg Endosc 2007;21(2):214–9.
[16] van Dongen KW. Distributed versus massed training:efficiency of training psychomotor skills. Surg Tech Dev 2011;1:40–2.
[17] Brunner WC, Korndorffer Jr. JR, Sierra R, Massarweh NN, Dunne JB, Yau CL, et al. Laparoscopic virtual reality training: are 30 repetitions enough? J Surg Res 2004;122(2):150–6.