# Performance Variables and Professional Experience in Simulated Laparoscopy: A Two-Group Learning Curve Study

*Jan-Maarten Luursema, PhD,*[*,†] Maroeska M. Rovers, PhD,[†,‡] Marleen Groenier, PhD,[§] and Harry van Goor, PhD*[*]

[*]Department of Surgery, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [†]Department of Operating Rooms, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands; [‡]Department of Epidemiology, Biostatistics & HTA, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands; and [§]Department of Technical Medicine, University of Twente, Enschede, The Netherlands

**OBJECTIVE:** Virtual reality simulators are increasingly used in laparoscopy training. Such simulators allow objective assessment of performance. However, both low-level variables and overall scores generated by the simulator can be hard to interpret. We present a method to generate intermediate performance variables and show how the resulting variables can be used to investigate the development of laparoscopic skills.

**DESIGN:** A beginner group (n = 16) and a group with intermediate laparoscopic experience (n = 9) participated in a 5-session, basic skills training course hosted by the Department of Technical Medicine at the University of Twente. Multiple simulator-generated variables were aggregated into 4 performance variables: duration, left-hand motion, right-hand motion, and damage. Differences in performance were analyzed in relation to proficiency values.

**RESULTS:** Damage performance differentiated the most between groups and proficiency values; motion performance variables differentiated the least. The more experienced group outperformed the beginner group at damage by the end of the course.

**CONCLUSIONS:** Differentiating between duration, left-hand motion, right-hand motion, and damage is a useful way to investigate laparoscopic performance development. Different performance variables follow different trajectories toward expertise. Valid and reliable clinical damage parameters are needed to investigate the relation of real-world damage to simulator damage. (J Surg 71:568-573. © 2014 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

**KEY WORDS:** laparoscopy, VR simulator training, performance assessment, damage control

**COMPETENCIES:** Practice-Based Learning and Improvement

*Correspondence*: Inquiries to Jan-Maarten Luursema, PhD, Department of Surgery, Radboud University Nijmegen Medical Centre, P.O. Box 9101, Internal Mail Code 960, Nijmegen 6500 HB, The Netherlands; fax: +31 2436-540501; e-mail: jan-maarten.luursema@radboudumc.nl

## INTRODUCTION

Virtual reality (VR) simulators are increasingly used to train the skills and procedures associated with minimally invasive techniques such as laparoscopy.[1,2] One of the main advantages of such simulators is the option to assess trainee performance quantitatively, allowing specific, objective trainee feedback and evaluation. To maximize VR simulator training effectiveness, trainee performance needs to be properly contextualized. This can be done by providing the trainee with valid and reliable, expert-based proficiency values, which allows competency-based training.[3,4]

Hindering the usefulness of VR simulator performance quantification is the following: VR simulators typically generate 2 types of performance metrics, namely numerous low-level variables and a single overall score. Overall scores are usually supplied by the simulator software and generated with unpublished procedures of uncertain validity. Additionally, overall scores are hard to interpret because of possible performance trade-offs. For instance, during task execution, a student may focus on error control and accept a duration penalty, or vice versa. Low-level variables allow detailed, near real-time task execution adjustment but are less suitable for evaluating performance in relation to overall course goals. This article presents a method for generating

  1931-7204/$30.00
http://dx.doi.org/10.1016/j.jsurg.2013.12.005

concise and useful performance variables from a multitude of simulator-generated data and shows how this method improves our understanding of the development of surgical skill, and can inform training course design.

Based on data from a yearly laparoscopic basic skills training course at the department of Technical Medicine, University of Twente (the Netherlands), we combined multiple low-level simulator-generated variables into 4 high-level performance variables. These are duration, left-hand motion, right-hand motion, and damage. The exact procedure and rationale is described in the "Data Reduction" section. For these performance variables, we analyzed the development toward expert-based proficiency values (published by Koen van Dongen et al.[5]), for a beginner group and a group with intermediate experience. We expected the intermediate group to outperform the beginner group. We also expected both groups to match expert performance toward the end of the course (based on previous editions of this course). No expectations toward the differences in development between the 4 performance variables were formulated.

## METHODS

### Ethics Statement

Written informed consent was obtained from all participants, all data were analyzed anonymously, and the study could not lead to potential or actual harm. Under Dutch law, medical educational research is exempt from formal ethical review, provided such research does not involve questionnaires collecting sensitive personal data (Medical-Scientific Research on Human Subjects; http://wetten.overheid.nl/BWBR0009408/geldigheidsdatum_20-11-2012, in effect since March 26, 1998). The study presented here qualifies as such, and consequently no formal ethics review was sought.

### Participants

We used data from 2 groups of participants. The first group, TG-2010, consisted of 16 master students of the Robotics and Imaging track of the Department of Technical Medicine (University of Twente, the Netherlands) who completed a required basic laparoscopic skills course in 2010. They were between 21 and 23 years of age, with a mean age of 21.8 years; 9 were women. All but one were right-handed. All reported no previous laparoscopic experience and also no previous laparoscopic simulator experience.

The second group, Medical Spectrum Twente hospital, consisted of 9 medical specialists practicing at the Medical Spectrum Twente hospital (5 women); they had intermediate laparoscopic experience, ranging from 5 to 250 procedures, with a mean of 70 (combined values for assisted procedures and procedures performed with primary responsibility, data gathered by self-report). All reported no previous laparoscopic simulator experience. This group was aged between 26 and 53 years, with a mean age of 33.9 years, and consisted of 3 surgeons and 6 gynecologists. None of the participants in this group were left handed. The unequal group sizes reflect the limited availability of participants.

### Procedure

Both groups participated in the basic laparoscopic skills training course at Technical Medicine's Experimental Center for Technical Medicine, a course that consists of 5 individual weekly 90-minute sessions on 2 different laparoscopic simulators, the ProMIS augmented reality simulator (Haptica, Dublin, Ireland) and the LapSim virtual reality simulator (Surgical Science, Gothenburg, Sweden). Overall, 6 LapSim tasks were practiced in 3 difficulty settings.[5]

Participants were instructed individually in the use of the simulator on a physical/ergonomic level and in how to use the simulator software to track one's progress on the low-level variables supplied by the software. The relevance of these low-level variables was explained in an electronic document made available to all participants. This document also contained descriptions of all tasks to be trained during the course.

Participants were free to control the number of repetitions for each task during a session, provided each task was at least practiced once.

### Simulator Task Selection

The basic skills LapSim tasks used in this training course were as follows: instrument navigation, coordination, grasping, lifting and grasping, cutting, and clip applying. They were selected after consultation with 2 experienced surgeons associated with the Department of Technical Medicine to represent the skill set required to start training specific laparoscopic procedures. A description and demonstration of the selected tasks can be found at Surgical Science's website, https://www.surgical-science.com/portfolio/lapsim-basic-skills/. The following tasks were left out: (1) suturing for lack of fidelity, (2) camera navigation for being a simplified version of instrument navigation and coordination, (3) the precision and speed skills because they were deemed covered by the instrument navigation and grasping tasks, (4) handling intestines was thought to be beyond basic skills, and finally (5) the fine dissection task was found to suffer from reliability issues (an easy way to cheat and still gather full score was found for this task).

### Apparatus

We used the LapSim laparoscopic simulator, which has been validated amongst others by Eriksen and Grantcharov, and Woodrum et al.[6,7] Immersion's Virtual Laparoscopic Interface hardware was connected to a Pentium 4 CPU 3.00 GHz, 504 MB RAM computer running Windows XP. Visual feedback to the participants was provided by a 19-in thin-film-transistor monitor. Surgical Science's LapSim v.3.0.10 software package provided the exercises.

## Data Reduction

All collected data were anonymized before analysis. Owing to technical difficulties with the ProMIS simulator throughout the training course, only LapSim data were used for analysis. Of the LapSim data, only the first trials for all moderate difficulty tasks were selected. For each training task, the LapSim surgical simulator reports a large number of low-level variables. To compensate for differences in magnitude caused by different measurement units or different task demands or both, all low-level training data were transformed into z scores before aggregating them into 4 overarching performance variables. This was done as follows.

### From Raw Data to Z Scores

For every simulator-reported low-level variable, a corresponding expert performance–based proficiency mean and standard deviation value was available.[5] Z scores were calculated by substracting same-variable expert mean values from each data point, and dividing the resulting value by the same-variable expert standard deviation. The resulting normalized variables express trainee performance in relation to expert performance, with zero indicating trainee performance equal to expert mean performance and "1 − 1" indicating performance at 1 standard deviation above or below the expert mean. Better performance for all low-level variables was indicated by lower scores, and this was retained by the aforementioned transformation.

### Aggregating Transformed Low-Level Variables Into Overarching Performance Variables

Next, we classified the transformed low-level variables in categories reflecting duration, left-hand motion, right-hand motion, damage, procedural error, and goal failure (Appendix 1). Variables in execution error and goal failure did not differentiate sufficiently between participants and were not further analyzed. The same was true for 3 of the 14 damage variables (Appendix 1). To analyze the data at a higher aggregation level, the 4 performance variables of duration, left-hand motion, right-hand motion, and damage were derived from the remaining low-level variables in 2 steps: first, for each task, mean values for the transformed raw variables within a performance category were calculated. This resulted in task-specific performance variables for duration, left-hand motion, right-hand motion, and damage. Second, mean values across tasks for all task-specific performance variables were calculated, resulting in the overarching performance variables used in our analysis of left-hand motion, right-hand motion, damage, and duration.

## Statistical Analysis

The development over a 5-session training course of the performance variables of left-hand motion, right-hand motion, duration, and damage was evaluated with respect to published expert-based proficiency values,[5] for 2 groups of different professional experience (beginner and intermediate). All variables of interest were normally distributed, as assessed by the Kolmogorov-Smirnov-1 test, allowing the use of standard inferential statistics. $T$ tests were used to compare means, with $\alpha$ set at 0.01 to correct for multiple testing. Cohen's $d$ is reported to indicate effect sizes. All analyses were performed using SPSS 18 (IBM/SPSS Inc, Chicago, IL).

# RESULTS

First, learning was established for all 4 performance variables in both experience groups by paired $t$ tests that compared first session with fifth session performance. Effect sizes were large, ranging from 1.10 to 1.93 (Cohen's $d$). T values ranged between 4.07 and 5.99 (all at $p < 0.01$).

## Performance Compared With Proficiency Levels

One-sample $t$ tests were calculated to compare first session and fifth session performances to expert-based proficiency values, for both groups and all performance variables (Table).

At the start of the course, beginners performed below proficiency values for all performance variables. At the end of the course, beginner performance for left-hand and right-hand motion was equal to proficiency values, but performance for damage and duration still was worse. These differences showed large effect sizes, Cohen's $d$ between 1.08 and 2.30, and $t$ values between 4.31 and 9.21 at $p < 0.01$.

The intermediate group performed initially below proficiency levels for damage and duration (Cohen's $d$ large, between 1.76 and 1.94; $t$ values between 5.27 and 5.83 at $p < 0.01$), but not for either left-hand or right-hand motion. At the end of the course, all performance variables were statistically equal to expert values, with damage still showing the largest but nonsignificant difference (medium effect size Cohen's $d = 0.76$; $t = 2.29$ at $p = 0.05$).

## Between-Group Performance Comparison

To directly compare performance between the beginner and the intermediate groups, 8 two-sample $t$ tests were done for all performance variables at the first session and the fifth session. Damage differed between the groups at the fifth

**TABLE.** One-Sample t Tests to Compare Experience Group Performance to Expert Values. Performance is Split by Performance Variable and Experience Group. The First and Fifth Sessions are Analyzed

| Performance Variable | Cohen's *d* | *t(df)* | p |
|---|---|---|---|
| Beginner group (technical medicine, n = 16) | | | |
| Left 1 | 1.08 | 4.31 (15) | <0.01 |
| Left 5 | −0.44 | −1.75 (15) | 0.10 |
| Right 1 | 1.62 | 6.49 (15) | <0.01 |
| Right 5 | 0.40 | 1.59 (15) | 0.13 |
| Damage 1 | 2.30 | 9.21 (15) | <0.01 |
| Damage 5 | 1.50 | 5.99 (15) | <0.01 |
| Duration 1 | 2.20 | 8.81 (15) | <0.01 |
| Duration 5 | 1.32 | 5.26 (15) | <0.01 |
| Intermediate group (Medical Spectrum Twente hospital, n = 9) | | | |
| Left 1 | 0.67 | 2.00 (8) | 0.08 |
| Left 5 | −0.40 | −1.20 (8) | 0.27 |
| Right 1 | 0.83 | 2.50 (8) | 0.04 |
| Right 5 | −0.45 | −1.34 (8) | 0.22 |
| Damage 1 | 1.94 | 5.83 (8) | <0.01 |
| Damage 5 | 0.76 | 2.29 (8) | 0.05 |
| Duration 1 | 1.76 | 5.27 (8) | <0.01 |
| Duration 5 | 0.48 | 1.44 (8) | 0.1 |

session, with the intermediate group outperforming the beginners (with large effect sizes for damage; $d = 1.19$; $t(23) = 2.78$ at $p = 0.01$).

### Training Volume

As participants were free to control the number of repetitions for each task (after a minimum of 1), the total number of trials could potentially confound performance as the course progressed. A 2-sample $t$ test for total number of trials showed no differences between the groups, confirming differences in performance between both experience groups and experts depend on professional experience alone (small effect size at $d = 0.35$; $t(23) = 0.827$ at $p = 0.42$).

## DISCUSSION

Different performance aspects follow a different trajectory toward expertise. Rather than treating proficiency as a single standard, we need to explore the consequences of this for the learning process and training course design. For both the beginner and the intermediate group, motion proficiency

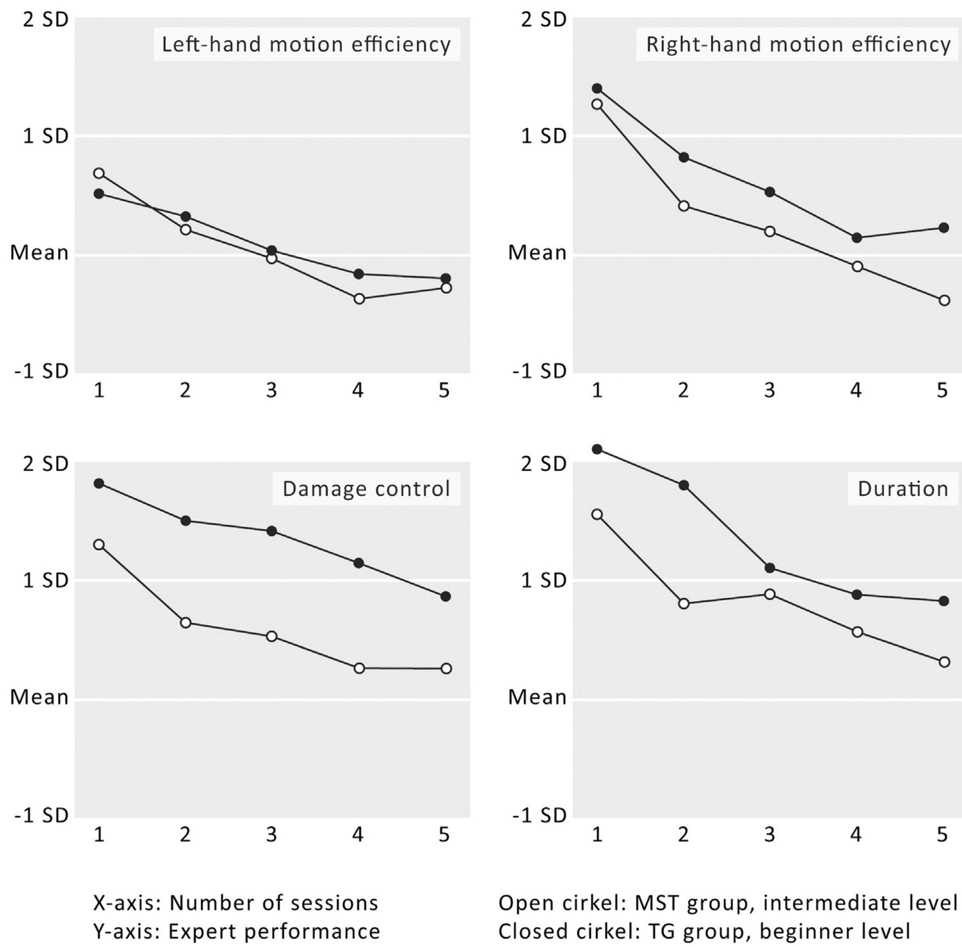


**FIGURE.** Learning curves for both groups and all 4 performance variables used in this study.

levels are reached before proficiency in duration and damage is reached (Fig.). For damage, we think this indicates a dissociation between damage and the other performance variables and points toward multiple parallel learning processes. In terms of theories of skills learning,[8] the primary task (motion control) is automated, freeing the residents' mental resources for secondary task performance (damage control). Support for this comes from Stefanidis et al.,[9] who found a similar pattern in a dedicated dual-task study, where the secondary task (a "visuospatial task") was not content related to the primary task (laparoscopic suturing). Thus, better damage control performance may indicate the extent motion performance is automated. As automated motion performance frees up mental resources for other task aspects, increasing damage control may be a better indicator than motion control for a student's readiness to start in the operating room.

Although the development of the duration variable was closer to damage than to either motion variable, duration may be better described as a second-order performance measure, being dependent on such different sources as damage, motion efficiency, and motion optimization. Thus, although duration is a clinically important performance measure (operating room cost and patient outcomes), improvements in duration do not tell us much about a person's learning phase. Because during late learning automation of skills reduces cognitive load,[10] being able to infer someone's learning phase would be essential to future work in predicting emergency performance in the operating room based on previous learning.[11] More work on nonduration performance measures is needed, to be able to both quantitatively assess operating room performance andrelate operating room performance to simulator training measures.

## LIMITATIONS

The extra training provided by the ProMIS simulator and the additional LapSim tasks in other than "moderate" difficulty settings may have sped learning in our students, which could have resulted in somewhat atypical, compressed learning curves. As this extra variable was the same for all students, we believe this does not affect our analysis or conclusions. Additionally, transfer from other (training) experiences to simulator performance may be typical but underreported for this type of study.

Goal failure and execution error measures, though in principle relevant for performance in the operating room, were not sufficiently differentiated in the current study to warrant analysis. This was mostly due to floor effects and ceiling effects. Customizing simulator settings may help increase the reliability of these measures for future studies. However, as both goals and procedures of basic skills tasks are very different compared with operating room goals and procedures, the relevance of basic skills goal failure and execution error measures for the operating room may be low.

## CONCLUSIONS

Simulator-recorded damage was shown in this study to be the variable that distinguishes best between different levels of laparoscopic expertise. Coupled with the clinical importance of damage control, this should render damage control a major focus of course design and of transfer studies. Work aimed at developing objective and quantified operating room performance measures for a range of laparoscopic procedures would be an essential first step to validate the resulting courses and to start research in the transfer of simulator damage control skills to the operating room.

## ACKNOWLEDGMENTS

Data collection for this research took place at the Department of Technical Medicine, University of Twente, the Netherlands. The authors thank Drs Heleen Miedema and Ir. Remke Burie of Technical Medicine for their cooperation and the students of technical Medicine as well as the medical specialists of the Medical Spectrum Twente Hospital for participating in this study.

## APPENDIX 1

Low-level variables reported by the LapSim laparoscopic simulator software are listed for the tasks used in the training course described in this study. Following the variable names listed are the abbreviated names for the tasks to which they apply; IN for Instrument Navigation, Coordination becomes CO, Grasping GR, Cutting CU, Clip Applying CA, and Lifting and Grasping becomes LG. Measures highlighted in bold type were not analyzed for lack of differentiation between participants (floor effects, ceiling effects, or otherwise insufficiently differentiated outcome values).

*Duration measures*:
Total time (s) (CO, CU, CA, LG)
Left instrument time (s) (IN, GR)
Right instrument time (s) (IN, GR)
*Left instrument motion measures*:
Left instrument path length (meters) (IN, CO, GR, CU, CA, LG)
Left instrument angular path (degree) (IN, CO, GR, CU, CA, LG)
*Right instrument motion measures*:
Right instrument path length (meters) (IN, CO, GR, CU, CA, LG)
Right instrument angular path (degree) (IN, CO, GR, CU, CA, LG)

*Goal failure measures*:
Misses (percentage of targets not met) (CO)
Left instrument misses (percentage of targets not met) (IN, GR, LG)
Right instrument misses (percentage of targets not met) (IN, GR, LG)
Timeout failure (percent) (CU)
*Execution error measures*:
Instrument outside view (number of times) (CO)
Instrument outside view (s) (CO)
Drop failure (percent) (CU)
Incomplete target areas (number of areas) (CA)
Badly placed clips (number of clips) (CA)
Dropped clips (number of clips) (CA)
*Damage measures*:
Tissue damage (number of times) (IN, CO, GR, CU, LG)
Maximum damage (millimeters) (IN, CO, GR, CU, LG)
Maximum stretch damage (percent) (CU, CA)
Rip failure (percent) (CU)
Blood loss (liters) (CA)

## REFERENCES


1. Aggarwal R, Balasundaram I, Darzi A. Training opportunities and the role of virtual reality simulation in acquisition of basic laparoscopic skills. *J Surg Res*. 2008;145(1):80-86.
2. Kneebone R. Simulation in surgical training: educational issues and practical implications. *Med Educ*. 2003;37(3):267-277.
3. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8): 638-645.
4. Leung WC. Competency based medical training: review. *Br Med J*. 2002;325(7366):693-696.
5. van Dongen KW, Ahlberg G, Bonavina L, Carter FJ, Grantcharov TP, Hyltander A, et al. European consensus on a competency-based virtual reality training program for basic endoscopic surgical psychomotor skills. *Surg Endosc*. 2011;25(1):166-171.
6. Eriksen JR, Grantcharov T. Objective assessment of laparoscopic skills using a virtual reality stimulator. *Surg Endosc*. 2005;19(9):1216-1219.
7. Woodrum DT, Andreatta PB, Yellamanchilli RK, Feryus L, Gauge PG, Minter RM. Construct validity of the LapSim laparoscopic surgical simulator. *Am J Surg*. 2006;191(1):28-32.
8. Fitts PM, Posner MI. Human Performance. Oxford, England: Brooks/Cole; 1967.
9. Stefanidis D, Scerbo MW, Korndorffer JR, Scott DJ. Redefining simulator proficiency using automaticity theory. *Am J Surg*. 2007;193(4):502-506.
10. Shiffrin RM, Schneider W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol Rev*. 1977;84(2):127-190.
11. Hsu KE, Man FY, Gizicki RA, Feldman LS, Fried GM. Experienced surgeons can do more than one thing at a time: effect of distraction on performance of a simple laparoscopic and cognitive task by experienced and novice surgeons. *Surg Endosc*. 2008;22(1): 196-201.