

Virtual Reality Simulation in Laparoscopic Gynaecology

PhD thesis

Christian Rifbjerg Larsen

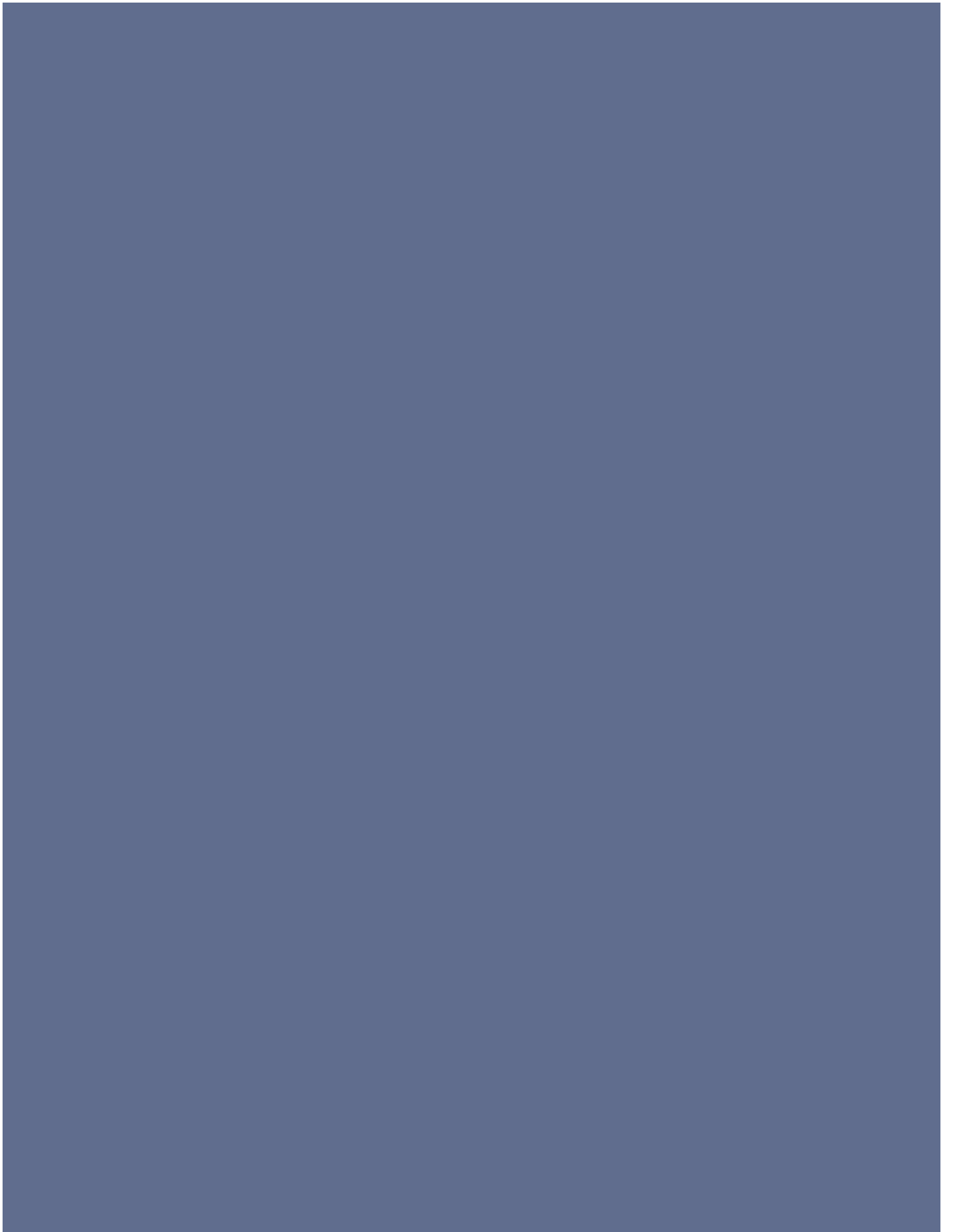
2009

Table of contents

4	Table of contents
7	Papers
8	Summary in English
10	Dansk resumé
12	Preface
13	Introduction
13	Hypotheses
13	Aim
14	Background
14	Historic development of surgical training
14	Laparoscopy: The new gold standard
14	Complications to laparoscopy
15	Reasons for complications
15	Special abilities needed for laparoscopy
17	Virtual Reality: Definition
17	Simulation: Definition
17	Virtual Reality Simulation in surgical skills training
19	Evidence for the use of laparoscopic simulators
19	Commercially Available Laparoscopic Simulators
20	Research strategy
21	Materials
21	Surgical procedure
21	Subjects and allocation
21	Paper I: Objective assessment of gynaecologic laparoscopic skills using the LapSimGyn virtual reality simulator.
21	Paper II: Objective assessment of surgical competence in gynaecological laparoscopy: Development and validation of a procedure-specific rating scale.
21	Paper III: Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial.
22	Simulator Equipment and Tasks
24	Results
24	Paper I: Objective assessment of gynaecologic laparoscopic skills using
24	Basic skills 1
24	Basic skills 2
24	Basic skills 3
25	Ectopic Pregnancy (Salpingectomy)

Table of contents

26	Paper II: Objective assessment of surgical competence in gynaecological laparoscopy:
	Development and validation of a procedure-specific rating scale.
26	Construct validity
26	Gamma coefficient
26	Kappa value
27	Paper III: Impact of Virtual Reality Training in Laparoscopic Surgery:
	A Randomised Controlled Trial.
27	Operative performance
28	Baseline characteristics and simulator training program
28	Inter-Rater Agreement & Gamma Coefficient
29	Discussion
29	Surgical competence
29	Simulation and the educational theory on surgical skills learning
30	Surgical Skills Improvement
31	Feedback on Surgical Skills Performances
32	Assessment of Technical Surgical Skills
33	Error analysis of Technical Skills
33	Time-Error matrix
33	Observational Clinical Human Reliability Assessment
34	Hybrid versions of the Time-Error matrix and OCHRA
34	Objective Structured Assessment of Technical Skills
36	Simulator Based Assessment of Technical Skills
36	Dexterity Analysis for Assessment of Technical Skills
37	Hierarchical Task Analysis of surgical procedures
37	Modified OSATS for Laparoscopic Salpingectomy (OSA-LS)
38	Principles for evaluating Assessment systems
38	Validity, Reliability, Inter-Rater Agreement, Kappa Statistic,
	Gamma coefficient & Kendall's Tau
38	Validity
38	Reliability
39	Inter-Rater Agreement & Kappa statistics
39	Gamma Coefficient & Kendall's Tau
40	Predictive validity
40	Transfer of skills
44	Certification
45	Perspectives and Future Studies
46	Acknowledgements
48	Reference list
54	Appendix



Papers

This PhD thesis is based on the following papers:

- I. Larsen C R, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, Ottesen B.
Objective assessment of gynaecologic laparoscopic skills using the LapSimGyn virtual reality simulator.
Surg Endosc. 2006 Sep;20(9):1460-6.
- II. Larsen C R, Grantcharov T, Schouenborg L, Ottosen C, Sorensen J L and Ottesen B.
Objective assessment of surgical competence in gynaecological laparoscopy:
Development and validation of a procedure-specific rating scale.
BJOG. 2008 Jun;115(7):908-16.
- III. Larsen C R, Sorensen J L, Grantcharov T, Ottosen C, Schouenborg L, Dalsgaard T, Schroeder T V and Ottesen B.
Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial.
Accepted, BMJ, 2009

The papers are referenced by their Roman numbers. Previously published papers will be reproduced in this text by permission of Springer-Verlag inc., New York, (Surgical Endoscopy), Blackwell Publishing, Inc. (British Journal of Obstetrics and Gynaecology) and BMJ Publishing Group Ltd (British Medical Journal)

This PhD thesis was submitted August 15, 2008 and accepted for defence November 10, 2008.

The defence took place January 16, 2009 at Rigshospitalet, Copenhagen, Denmark.

Academic advisors: Professor DMSc Torben V. Schroeder, MD MEd Jette Led Sørensen, MD Ph.D. Teodor P Grantcharov and Professor DMSc managing director Bent S Ottesen

Official opponents: Professor DMSc Jacob Rosenberg (chair) Professor Ph.D., MHPE, MI. Berit Eika and DMSc Svend Schulze.

Disclosure: This thesis, including all papers, was made independently from commercial interests, and there is no conflict of interests. The work was financed by the Juliane Marie Centre, Copenhagen University Hospital, Rigshospitalet, The Research foundation of Rigshospitalet, and the following private donors: Tryg Foundation, Toyota Foundation, Alternative Foundation, Aase and Ejner Danielsens Foundation, and Classenske Fideicommis.

Summary in English

PhD Thesis: **Virtual Reality Simulation in Laparoscopic Gynaecology**

Aim

To investigate, on novice and experienced gynaecological surgeons:

1. The construct validity of the LapSim[®]Gyn Virtual reality Simulator, and to determine the learning curves of novice gynaecologists. To establish the expert performance level in the simulator.
2. To develop and validate a global and a procedure-specific rating scale for assessment of technical skills in laparoscopic gynaecology.
3. To investigate if skills obtained by simulator training can be transferred to human operation

Background

Laparoscopic surgery requires proficiency with sophisticated technical skills. Reports of complications caused by impaired technical surgical skills have highlighted the importance of teaching surgical technical skills in a safe, realistic and efficient environment. The traditional approach of 'see one do one, teach one' is no longer acceptable to either the surgical profession or to the well-informed and demanding patient. There is also a need for unbiased structured objective assessment of technical skills during the surgical education. Virtual Reality Simulators might possess the capacities needed for future basic training in laparoscopic surgery, however, there is little research evidence of their efficiency and little is known on the transferability of skills beyond the artificial environment of the setting of the training facility.

Research Strategy

For the investigation we choose the Virtual Reality Simulator LapSimGyn, in which both basic skills and complete operative procedures can be trained. The evaluation was executed in three stages:

1. Evaluation of the construct and discriminative validity of the simulator, generating learning curves for novice gynaecologists and determine the expert performance level in the simulator. Design: Prospective cohort.
2. Developing and validating a general and procedure specific rating scale for evaluation performance in laparoscopic gynaecology. Furthermore, to investigate the Inter-Rater Agreement, the gamma coefficient (Kendall's rank correlation) which is a measure of the strength of dependence between observations, and the Kappa value for each of the ten individual items included in the rating scale. Design: Prospective cohort, observer blinded study
3. Establishing the effect of procedural Virtual Reality Simulator training on a human laparoscopic operation. Training was criterion based and the novices in the intervention group had to train until they reached the expert performance level defined in the first study. Outcome was operative performance assessed by observers blinded to subject and group status, using the rating scale validated in the second study. Design: Prospective Randomised Controlled and blinded trial

Summary in English

Results

Data from the first study showed that expert gynaecologists performed significantly and consistently better than intermediate and novice gynaecologists. Learning curves differed significantly between the groups, showing that experts start at a higher level and more rapidly reach the plateau of their learning curve when compared to intermediate and novice groups of surgeons. The second study demonstrated significant differences in surgical performance between the three groups, hence the rating scale was both construct and discriminative valid. The Inter-rater agreement, kappa value and gamma coefficient was sufficiently high. Finally, in the intervention study, the simulator trained group reached a mean total score as intermediate experienced gynaecologists while the controls performed as true novices. The mean total operating time was reduced with 50% in the simulator trained group, both findings highly significant.

Conclusion

The LapSimGyn VR simulator demonstrates construct validity, both on Basic Skills module and on the procedural gynecological module, hence the simulator can be used for both training and assessment. The procedure-specific rating scale for laparoscopic salpingectomy is a valid and reliable tool for assessment of technical skills in gynecologic laparoscopy. Skills in laparoscopic surgery can be clinically relevant increased by proficiency based procedural virtual reality simulator training. The performance level of novices is increased to the level of intermediately experienced laparoscopists and the operation time is reduced substantially. Mandatory simulator training should be considered before trainees perform laparoscopically on humans.

Ph.d. Afhandling: **Virtual Reality Simulatortræning i laparoskopisk gynækologi.**

Formål

Formålet med undersøgelsen var at vurdere om en computersimulator til operationstræning kan anvendes:

1. til at skelne objektivt mellem forskellige færdighedsniveauer i laparoskopisk kirurgi
2. som træningsværktøj i indøvelse af operative færdigheder
3. om færdigheder opnået ved simulatortræning kan overføres til operationer på patienter

Baggrund

Kikkertkirurgi anvendes stadig oftere i Danmark. Hidtil har træningen af de kirurgiske færdigheder manglet systematik og tilstrækkelig evaluering af kirurgerne. Desuden foregår den basale træning stadig på patienter. Computerbaserede operationssimulatorer (virtual reality) kan være fremtidens trænings- og evalueringsværktøj i den kirurgiske uddannelse, ligesom flysimulatorer er det for piloter. Det er dog fortsat uafklaret hvorvidt simulatorerne er valide som evalueringsværktøj, og om færdigheder opnået ved træning i simulatorer kan overføres til egentlige operationer. Det vil have en stor klinisk og uddannelsesmæssig betydning, hvis computerbaserede operationssimulatorer er et værktøj, der giver mulighed for en realistisk optræning af operatøren til et niveau der optimerer den kliniske uddannelse på operationsstuerne. I givet fald vil det minimere risici og komplikationer der skyldes manglende øvelse, samtidig med at simulatoren kan anvendes som løbende evalueringsværktøj for de uddannelsessøgende.

Forskningsstrategi

LapSimGyn simulatoren blev valgt fordi den både giver mulighed for at træne basale færdigheder og hele operationsforløb. Valideringen af simulatoren blev opdelt i tre faser:

1. Prospektivt kohortestudie til fastlæggelse af simulatorens konstruktive og diskriminative validitet. Den konstruktive validitet beskriver hvorvidt det indbyggede evalueringssystem tester det vi ønsker at teste: Laparoskopisk tekniske færdigheder. Den diskriminative validitet angiver hvorvidt simulatoren kan skelne mellem forskellige færdighedsniveauer. Desuden blev læringskurver for novicer samt referenceniveauer for eksperter kortlagt.
2. Prospektivt kohortestudie til at udvikle og validere en skala til evaluering af laparoskopisk kirurgisk færdigheder hos læger med forskellig klinisk operativ erfaring. Bestemmelse af konstruktiv og diskriminativ validitet, inter-observatørvarians, kap-paværdi, gamma koefficienter for skalaen var alle elementer der indgik i vurderingen.
3. Et prospektivt randomiseret kontrolleret forsøg til at undersøge effekten af forudgående simulatortræning på operationer på patienter. Interventionsgruppens deltagere gennemførte simulatortræning indtil de hver især nåede færdighedsniveauet defineret af eksperterne i det førstedelstudium. Kontrolgruppen gennemførte den traditionelle uddannelse som assistenter på operationsstuerne. Alle gennemførte efterfølgende en laparoskopisk salpingektomi der blev videofilmet. Operationerne blev evalueret af to uafhængige eksperter blindet for forsøgsdeltager og gruppestatus.

Resultater

1. Simulatoren er valid som evalueringsværktøj for uddannelsessøgende. Simulatoren's evaluerings-system er i stand til at diskriminere mellem forskellige erfaringsniveauer for kirurger. Eksperterne kan sætte et referenceniveau som novicerne kan arbejde op imod. Novicernes læringskurve kan bestemmes løbende.
2. En skala til systematisk objektiv evaluering af laparoskopiske salpingectomier på patienter blev udarbejdet. Skalaen viste sig i en separat prospektiv undersøgelse både konstruktiv og diskriminativ valid. Evalueringsskalaen har en lav inter-observatør variation, en høj gamma koefficient samt en høj kappa værdi.
3. Simulatortræning har en dramatisk effekt på novicernes operative færdigheder. De trænede novicer scorede det samme som mellemøvede operatører (erfaring fra 30-50 operationer) medens kontrolgruppen scorede som sande novicer. Den simulator-trænede gruppe brugte 50 % mindre tid på at gennemføre den samme operative procedure.

Konklusion

Virtual Reality simulator træning bør indføres som obligatorisk træning før yngre læger opererer laparoskopisk på patienter. De opnåede færdigheder bør certificeres. Den uddannelsessøgendes kirurgiske udvikling bør følges og understøttes ved anvendelse af validerede skalaer for laparoskopisk kirurgiske færdigheder. Med disse skalaer kan den uddannelsessøgende få løbende struktureret objektiv feedback og få monitoreret de nødvendige færdighedsmæssige fremskridt.

Preface

“The scientific attitude lies at the heart of the scholarship and is accepted by everyone in the field.

The situation seems quite different in education. We do the things we do, because that is the way we have been raised ourselves and that is the way it has been done for many years, even centuries. We hardly read the literature on education, or, more appropriately, are not even aware that such literature exists.

It is difficult to change things in education, because as teachers we are highly convinced that what we do is appropriate and any challenge to one’s convictions is an actual challenge to one’s professional integrity”¹.

CPM Van Der Vleuten

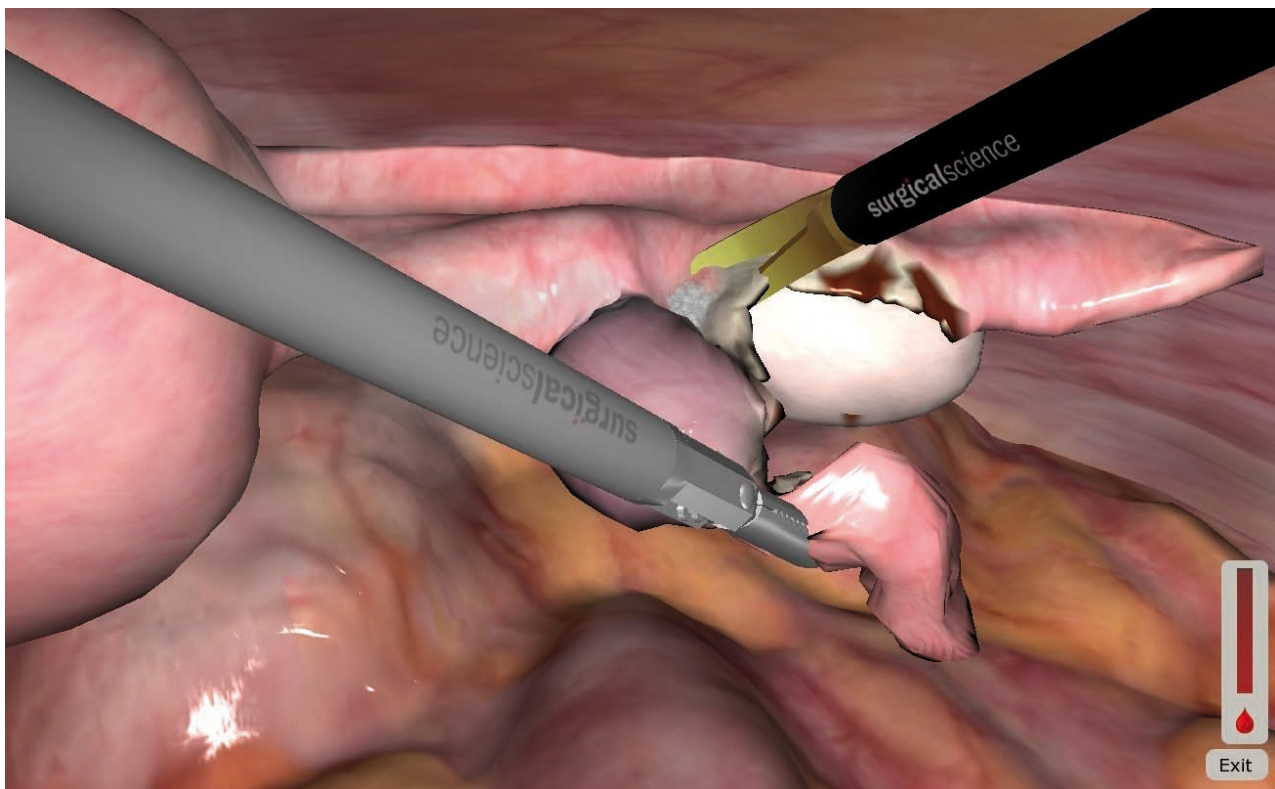


Figure 1: LapSimGyn Virtual Reality Simulator: Salpingectomi

Introduction

During the past decades there have been growing professional, public and health authority concern over patient safety and rights, using patients as subjects for surgical training, medical costs, complications to surgery³ and about the quality of the postgraduate surgical education⁴. The traditional surgical education has been based on the apprenticeship model, unstructured and limited by the supply and not the demand for training procedures. At the same time laparoscopy has become the new gold standard surgical approach in an increasing number of procedures. The psychomotor and perceptual obstacles for learning laparoscopic surgery, and the fact that the operation room is a stressful and cost-intensive facility for the training of basic skills have put additional pressure on the traditional educational model. Reduced work time as well as demands for

structured, criterion-based education further increases the demand for new educational strategies in surgery⁵. Simulation in healthcare education has been known for nearly 40 years, and during the past decade we have witnessed widespread adoption of the technology⁶. Thanks to general accessibility to cheaper high-performance computers it has become possible to develop and implement computer-based virtual reality systems in medical education. Virtual reality simulators for surgical training might possess the properties needed for future basic training in laparoscopy. However, there is little scientific evidence of simulators' efficiency and little is known of the transferability of skills beyond the artificial environment of the training facility.⁷

Hypotheses

The Hypotheses for this thesis was to investigate if:

1. The virtual reality simulator LapSimGyn can be used for a reliable and valid assessment of technical surgical skills, thereby discriminating between different levels of laparoscopic proficiency.
2. Learning curves can be generated in the LapSimGyn virtual reality simulator, monitoring the improvement of technical skills during simulator training.
3. Criterion-based virtual reality simulator training courses can increase the surgical performance of novice gynaecological laparoscopists, thereby moving the basic skills training from the patients in the operating rooms to a safe, less stressful and less cost-intensive environment

Aim

The aim for this thesis was to investigate the validity of the LapSimGyn virtual reality simulator for basic training and assessment of laparoscopic skills, according to the hypothesis listed. The aim was divided into three topics elucidated separately:

1. The validity of the simulator as a tool for the assessment of technical laparoscopic skills
2. The development and validation of a procedure-specific, objective and structured rating scale for the assessment of surgical skills in laparoscopic gynaecology.
3. The validity of the simulator as an instrument for training laparoscopic surgical skills, measured as impact of simulator training on operative performance in gynaecological laparoscopy.

Background

Historic development of surgical training

During the past century surgical training has been based on an apprenticeship model characterised by the paraphrase “see one, do one, teach one” introduced by the surgeon William Stuart Halsted (1852-1922) in his work “The Training of the Surgeon” from 1904⁸.

This approach to training is based upon the assumption that expert level is reached just through extensive experience, an assumption which lacks substantial empirical evidence⁹. Though the concept is challenged today it has produced excellent surgeons over time. More recently portfolios, check-list and log books have been introduced and the educational goals have been redefined turning away from sheer quantity of performed procedures and towards quality or competence level of the performed procedures⁴. Today accreditation is central for benchmarking clinical work and standards, most likely becoming a key issue in education. The demand for certified continuous medical education (CME) will also intensify, thus increasing the demand for a quality- rather than quantitative-based approach to education and training. During the past two decades there has been a substantial shift towards an evidence-based approach in clinical medicine. This shift is now also emerging in the field of education, training and assessment of medical professionals.

Laparoscopy: The new gold standard

Minimally invasive techniques have become the gold standard in many surgical fields during the past 20 years. This development has been driven by the advantages of lesser surgical trauma, faster post-operative recovery, shorter hospitalisation, better cosmetic results and a desire for sales by the medical industry. In the 1970s, laparoscopy in gynaecology was restricted to simple procedures such as sterilisations and diagnostic laparoscopy. After the first gynaecological operations had been performed laparoscopically, a salpingo-oophorectomy in 1987¹⁰ and a laparoscopically assisted

vaginal hysterectomy (LAVH) in 1988¹¹ the approach has become first choice in a majority of benign gynaecological conditions and supplementary to open surgery in some malignant cases. In Denmark (5.6 million inhabitants) in 2002, 94% of all surgically treated ectopic pregnancies were managed laparoscopically¹². Also more advanced laparoscopic procedures like retro-peritoneal lymph node dissection in malignant diseases is becoming more common.

Complications to laparoscopy

However, after the first enthusiasm, it was clear that the laparoscopic technique was associated with an increased number of complications and longer operation time at least during the initial part of the learning curve. This has been verified in many different fields e.g. general^{13,14} urological^{15,16}, paediatric¹⁷ and gynaecological surgery¹⁸. Unfortunately most authors reporting in this field fail to distinguish between resident-level trainees, who by definition are supervised (or should be) and attending surgeons who are introducing a new procedure to their practice and most often are unsupervised early in their learning curve. Although operative time might be longer with novice resident trainees, the final outcomes of a supervised operation should not be different. This is in distinct contrast to attending surgeons who might, after a short course in a new technique, operate unsupervised on their first patient the following week.

Background

Reasons for complications

Papers on complications in laparoscopy mostly deal with the frequencies of different injuries, and hardly ever analyse the reasons for the complications. In a paper based on data from The Physicians' Insurers Association of America, the most common reasons for laparoscopic surgical injuries leading to liability claims were analysed¹⁹. The top three reasons were: inexperience and defects in eye-hand coordination leading to a long learning curve, insufficient knowledge of human three-dimensional anatomy, and lack of knowledge of the equipment and control of the instrument's functionality. It has also been documented that it is possible to overcome these learning problems by appropriate training and by ensuring that the individual surgeon performs a sufficient quantity of procedures²⁰. However these findings are possible but hardly comforting to the patients having surgery performed during the early stage of a surgeon's learning curve.

Special abilities needed for laparoscopy

Research has shown that the speed of acquisition of laparoscopic technical skills does not differ between novice surgeons and experts in open surgery²¹. As the laparoscopic technique was put into practice by experts in open surgery, they were suddenly, on a technical level, turned into novices. It was evident that laparoscopy has a characteristic learning curve²² and that the traditional model for training surgical skills was insufficient and had to be supplemented by new training methods²³. Compared to open surgery, laparoscopic surgery possesses distinct features making learning a challenge on several levels. The long learning curve in laparoscopy is caused by the fundamentally different skills demanded compared to skills demanded in open surgery. The primary obstacles in learning laparoscopy are of a psychomotor and perceptual nature.

The psychomotor challenge stems from the long instruments used together with the effect of the body wall displacing the axis on which the instruments are moved around, which causes the tip of the instrument to move in the opposite direction to the surgeons hand, the fulcrum effect or "counterintuitive hand movements"^{24,25}. Operation through ports (trocars) also reduces the manoeuvrability of the instruments compared to open surgery²⁶.

The perceptual challenge is of both visual and tactile nature. Visually the surgeon has to navigate the instruments in a three-dimensional room using only two-dimensional visual feedback, on a small and distant video monitor²⁷. Another perceptual challenge is the vague haptic feedback using laparoscopic instruments. The surgeon is almost solely dependent on the impaired visual feedback.

These obstacles to learning laparoscopic surgery emphasise the need for a novel approach to training and assessment of skills in laparoscopy. During the past decade numerous commercially available training systems have been developed and marketed. Physical training models like Box- or Video Trainers, computer based models like Virtual Reality Simulators and Hybrid Simulators combining Box-trainers and computer based systems have flooded the market. An overview of various types of training systems and the level of evidence for their use has been listed in table 1. Nevertheless, despite the overwhelming common-sense acceptance of their importance in medical education, the educational benefit of clinical skills simulation remains relatively unproven²⁸.

Table 1
Commercially available Simulators for laparoscopic surgery, adopted from Ugeskrift for Læger 2006²⁹

Simulator (Manufacturer)	Price (Euro)	Type	Mobil / stationary	Force Feedback	Evaluating system	Basic skills module	Procedural module(s)	Evidence for assessment (construct / contrast validity)	Evidence for training (Predictive / concurrent validity)
SimSurgery 1. Basic skills 2. Lap Cholecystectomy 3. Gynaecology 4. Robotic surgery (SimSurgery)	33,000	Virtual Reality Simulator	Stationary	No (accessory)	Metrics values on instrument movements Time Error parameters	Camera navigation abstract and anatomical modules and suturing	Lap. Cholecystectomy Gyn. Module: -Ectopic pregnancy -da Vinci robotic surgery modules	No studies published	No studies published
LapSim 1. Basic skills 2. Cholecystectomy 3. Gynaecology (Surgical Science)	40,000 ¹	Virtual Reality Simulator	Stationary / mobile	No (accessory)	Metrics values on instrument movements Time Error parameters	Camera navigation abstract and anatomical modules and suturing	Lap. Cholecystectomy Gyn-module: -Ectopic pregnancy -Uterine Myoma -Sterilisation	2b 2b 2b	2b 1b Cholecystectomy 1b (Submitted)
Lap Mentor 1. Basic skills 2. Cholecystectomy 3. Lap. Gastric by-pass (Simbionix)	130,000	Virtual Reality Simulator	Stationary	Yes (electronic, virtual instruments)	Metrics values on instrument movements Time Error parameters	Camera navigation Abstract models and suturing	Lap. Cholecystectomy Hernia repair	4 4 No studies published	No studies published No studies published No studies published
Procedicus MIST 1. Core skills I og II 2. Intrakorp. suturing (Mentice)	37,000	Virtual Reality Simulator	Stationary / mobile Time Error parameters	No (accessory)	Metrics values on instrument movements	Abstract models and suturing	No	2b No studies published	2b No studies published
ProMIS (Haptica)	53,000	Hybrid: Virtual Reality and Box	Stationary	Yes, mechanical, real instruments	Box-trainer with computer registration of movements and time	Abstract models Standard instruments Diathermy included	No	2c	No studies published
Lap Trainer (SIMULAB)	1,300	Box-trainer	Stationary	Yes, mechanical, real instruments	No, depends on a supervisor	Abstract models Standard instruments Diathermy optional	No, but high fidelity models	No studies published	No studies published
LaparoTrainer (LINA)	4,000	Box-trainer real instruments	Portable	Yes, mechanical, Standard instruments Not diathermy	No, depends on a supervisor	Abstract models	No	No studies published	No studies published
TRSPN05 (3-D med)	4,000	Box-trainer real instruments	Stationary	Yes, mechanical, Standard instruments Not diathermy	No, depends on a supervisor	Abstract models	No	No studies published	No studies published
Blurup-Jensen DIY model	500	Box-trainer	Portable	Yes, mechanical real instruments	No, depends on a supervisor	Abstract models Standard instruments	No	No studies published Not diathermy	No studies published

Background

Virtual Reality: Definition

Virtual reality can be defined as a technology which allows a user to interact with a computer-simulated environment, be it a real or imagined one. Most current virtual reality environments are primarily visual experiences, displayed on a computer screen, but some simulations include additional sensory information, such as sound or haptic or tactile information, generally known as force feedback, in medical as well as aero-space and game applications. The simulated environment can be similar to the real world, for example, in simulations for aero-space, naval and medical personnel or it can differ significantly from reality, as in most video games. In medical simulators the environment can be completely abstract as seen in the first generations of simulators or mimic biological tissues as in second generation simulators (low fidelity) or mimic human anatomy as seen in the third generation of simulators (high fidelity). In the present study we chose a high fidelity simulator in order to simulate a complete surgical procedure in a life-like environment.

Simulation: Definition

Different definitions of simulation have been presented in the literature³⁰⁻³⁴. In a review by Issenberg et al., one of the definitions is broad and seems to cover many of the others: "In broad simple terms a simulation is a person, device or set of conditions which attempts to present education and evaluation problems authentically. The student or trainee is required to respond to the problems as he or she would under natural circumstances. Frequently the trainees receive performance feedback as if he or she were in the real situation"³¹. This definition covers the LapSimGyn virtual reality simulator used in the present study.

Virtual Reality Simulation in surgical skills training

Simulation in medical training and education may have an impact on, among other things,, risk management, life-long learning, education, training and continuous personal and professional development, staff management, continuous quality improvement, assessment and certification as well as management of poor performance.^{31,35} The expert-opinion-based advantages of simulators including my comments are listed in table 2. Despite the common sense acceptance of simulators' impact on skills improvement, evidence in the area of medical simulation is still sparse.^{28;30;35-37}

Background

Table 2

Current expert opinions on medical simulators³⁵, with comments by CR Larsen, and with suggestions for further research

Possible advantages of simulation	Piece of information	Comment	Research / development
1 Risks to patients and learners are avoided	Established	Simulators are safe environment and have no contact to the patients	Not necessary
2 The training agenda is determined by the needs of the learner, not the patient	Established	Simulation is independent of patients	Not necessary
3 Environment is safe. Learners have "permission to fail" and learn from such failure in a way that would be unthinkable in a clinical setting	Established	Simulators are safe and independent of patients	Not necessary
4 Undesired interference is reduced	Established	Simulator training can be executed at suitable time and place	Not necessary
5 Tasks/scenarios can be created according to demand	Theoretically true	Depends on needs and technical and creative abilities	Continuously development of simulated environments
6 Skills can be practised repeatedly	Established	Fundamental principle of simulation	Not necessary
7 Learner can be focused on the whole procedures or specific components of a procedure	Established	Depends on level of complexity	Not necessary
8 Training can be tailored to individuals	Established	Depends on level of complexity	Continuously development of simulated environments
9 Retention and accuracy are increased	Unknown	Lack of quality studies	Prospective studies are needed
10 The simulators offer the potential of providing feedback both for collaborative and for	Partly Established	Especially simulators for anaesthesia are designed to provide both collaborative and individual training/feedback individual learning	Further studies especially regarding surgical are needed
11 Transfer of training from the teaching situation to a real life clinical situation is enhanced	Unknown	Insufficient data	Randomised controlled trials is needed
12 Simulators can provide objective evidence of performance	Partly Established	More and more simulators are investigated for discriminative validity	Validation studies must be carried out before relying on new types of simulators
13 Simulators can be used in certification	Unknown	Depends on construct validity and defined criterion levels	Further development and validation of objective test must be provided

Background

Evidence for the use of laparoscopic simulators

Until now investigations of the effect of simulator training have primarily been carried out in first generation simulators and the results are ambiguous. A few studies on basic skills training have suggested an effect³⁸⁻⁴², however none of them exceed evidence level 2a, and a single study on procedural training of laparoscopic cholecystectomy has indicated a positive effect of simulator training at a 1b evidence level⁴³. One study demonstrated no positive effect of simulator based training⁴⁴, and one study was conducted by the manufacturer.⁴⁵ However, only few studies possess sufficient quality to reach evidence level higher than 2, and so far no Cochrane meta-analysis has been conducted. Until now no research has been published in the area of gynaecology (Table 3): However, the need for better and more structured laparoscopic surgical education is widely accepted. So far, there has been a tradition for the application of rigorous scientific methods in medical research but the existence of a remarkable difference in attitude between university staff in their role as researchers and that as teachers is noteworthy¹. The largest reviews so far on simulator-based training and assessment substantiates this notion, by concluding that the research in this educational area is sparse and, except for a few pioneering studies like those by Grantcharov et al.³⁸ and Seymour et al.³⁹, of inferior quality compared to clinical research^{31,37}.

Commercially Available Laparoscopic Simulators

Various laparoscopic simulators, mechanical or computer-based, are accessible for training the special skills needed for laparoscopic surgery. A virtual reality simulator is generically designed as a software program running on a computer which is connected to a user (surgeon) interface. The first simulators were simple, based on abstract models for manipulation in a virtual 3-dimensional room. The early abstract models primarily trained hand-eye coordination and dexterity. Due to the fast-increasing development of computer power, more and more sophisticated software programs were developed, typically with anatomically-correct, high fidelity body imaging offering realistic training. The latest models offer complete procedural training of for instance cholecystectomy, hernia repair or various gynaecological operations, thereby also training procedural knowledge. The hand-piece mimics surgical instruments and serves as an interface between the surgeon and computer software. The interface can have tactile feedback, however this multiplies the price, and the utility value is not evident.

The latest virtual reality simulators from LapSim, SimSurgery and Simbionix offer both a combination of basic skills training, dexterity, hand-eye coordination, instrument handling and training of complete operative procedures in the procedural modules, thereby also training cognitive functions such as knowledge of operative procedures, as well as decision making ability²⁹. A list covering the most used simulator systems has been collected in table 1.

In general, as noted above, only low to intermediate grade evidence exists regarding the use of simulators both as instruments for assessment and training of technical surgical skills. There is so far no evidence for the use of simulators in gynaecology. The present work has been performed in order to remedy this defect, by validating a Virtual Reality Simulator for gynaecological laparoscopy.

Research strategy

It was decided to conduct a three-step investigation to meet the aims.

First step

Validation of the LapSimGyn virtual reality simulator's assessment system and definition of expert criterion level, using the laparoscopic salpingectomy module. Obtain learning curves on simulated salpingectomy for novice, intermediate experienced and experienced laparoscopists. This was carried out by a prospective cohort study: (Paper I)

Second Step

Through a Hierarchical Task Analysis and comparative analysis of laparoscopic salpingectomy to develop an objective and structured rating scale for human laparoscopic salpingectomy. Investigate the construct and discriminative validity and estimate the inter-rater agreement, the gamma correlation coefficient and the Kappa value of the developed rating scale. The investigation of the rating scale was carried out as a prospective cohort study: (Paper II)

Third step

Predictive validity of the simulator. In this step the transferability of skills obtained by simulator training to a human laparoscopic salpingectomy was investigated. Trainees in obstetrics and gynaecology were included in a prospective, randomised, controlled and blinded trial. The intervention group undertook simulator training until they reached the expert criterion level based on expert performance defined in the first research step. The control group undertook traditional clinical education (assistance and camera navigation in the operating room). The outcome was established using the previously validated rating scale "Objective Structured Assessment of Laparoscopic Salpingectomy" from the second step investigation. The assessment was carried out by two independent observers blinded to subject and group status. (Paper III)

Materials

Surgical procedure

The laparoscopic salpingectomy was chosen as an index operation for several reasons. First, laparoscopic salpingectomy is a key operation in gynaecology. It is used for the potentially life-threatening condition ectopic pregnancy, and should be mastered by all gynaecologists and obstetricians working in on call duty. Second, the procedure possesses the necessary complexity for the assessment of laparoscopic skills, and third, the complete procedure can be trained in the LapSimGyn simulator.

Subjects and allocation

Paper I: *Objective assessment of gynaecologic laparoscopic skills using the LapSimGyn virtual reality simulator.*

The participants were a cohort (n=30) of gynaecological trainees, junior- and senior consultants, grouped according to their laparoscopic experience. We had no gold standard for the assessment of laparoscopic competence; consequently participants were grouped according to their experience, not to be taken as competence. Surgical experience was defined as numbers of procedures performed at a given level of complexity, consequently not based on an objective judgment of skills⁴⁷. Novices had no experience from laparoscopic surgery; intermediate experienced had performed 30–60 procedures and experts more than 200 procedures.

Paper II: *Objective assessment of surgical competence in gynaecological laparoscopy: Development and validation of a procedure-specific rating scale.*

The validation of the rating scale (Assessment chart) developed for procedure-specific, objective, structured assessment of laparoscopic salpingectomy (OSA-LS) was based on the evaluation of 21 randomly collected video recordings of laparoscopic salpingectomy. The procedures were performed by novices (n=6), intermediate experienced (n=7) and experts (n=8).

Paper III: *Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial.*

For the randomised controlled trial we invited all gynaecological trainees without previous experience in advanced laparoscopy, who worked in the region of Zeeland in Denmark. The first 24 eligible subjects (year 1–2 in specialist training) were randomised based on stratification according to experience in simple laparoscopy. Simple laparoscopy was defined as single instrument procedures like diagnostic laparoscopy and single instrument sterilisations. All procedures including coordination of two or more instruments were considered advanced laparoscopy. Randomisation was carried out in an external, computer-based and concealed procedure. (Intervention group n=13 Controls n=11). Two independent senior consultants assessed the laparoscopic performances blinded to subject and group status. The senior consultants are experts in laparoscopic surgery and have performed more than 2000 advanced laparoscopic operations.

Materials

Simulator Equipment and Tasks

The investigations were carried out using LapSim[®]Gyn v 3.0.1 Virtual Reality Laparoscopy Simulator (Surgical Science A/B, Gothenburg, Sweden). The program was executed on an IBM T42[®] Computer with a 15" colour monitor (Pentium M 1.8 GHz /512 MB RAM, IBM Inc., Armonk, NY, USA) running Windows XP professional[®] disc operating system (Microsoft, San Francisco CA, USA). The computer was linked to Virtual Laparoscopic Interface[®] with a diathermy pedal (Immersion inc., San Jose, CA, USA). The setup mimics the real gynaecological laparoscopic setup in which the surgeon is standing lateral to the patient and observing the monitor at eye level at the patients' feet, a distance of 1½ metres. The hand-piece was covered by surgical draping preventing direct visual-based navigation of the tip of the instrument. The hand-piece has five degrees of freedom and the interface unit transmits all instrument movements in real-time to the computer.

The LapSim Basic Skills module consists of 10 different tasks:

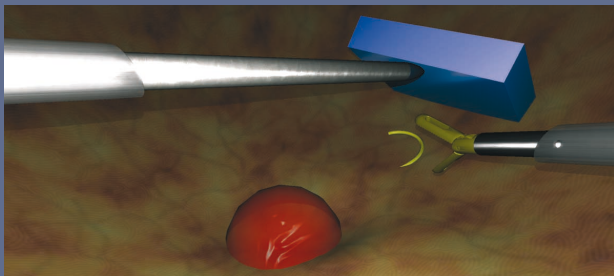
1. Camera navigation
2. Instrument navigation
3. Coordination
4. Grasping
5. Cutting
6. Clip applying
7. Lifting and Grasping
8. Suturing
9. Precision and Speed
10. Handling Intestines
11. Fine Dissection

The Gynaecological module consists of four tasks:

1. Myoma Suturing
2. Tubal Occlusion
3. Tubotomy
4. Salpingectomy

Figure 2

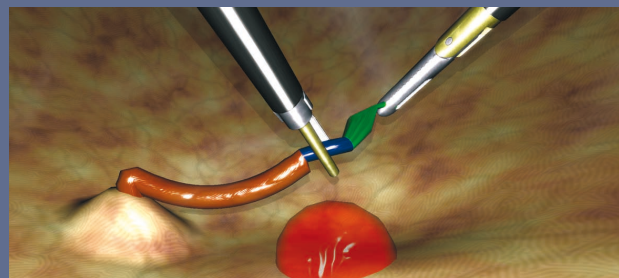
Graphic task presentation:
Lift and Grasp



The first basic skills program "Lift and Grasp" also trained handling of grasper and coordination between hands. It consisted of a symmetric left- and right-hand task in which a larger object should be lifted by one hand, releasing a small object that should be grasped gently by the other hand, moved and released in a defined target area. The task alternated between right and left hand making it possible to test and compare the surgeon's dominant versus non-dominant hand. cf. Figure 2

Figure 3

Graphic task presentation:
Cutting



The second task "Cutting" trained tissue handling, presentation of tissue and use of scissors. The surgeon must grasp a small, soft and sensible blood vessel with a grasper in the right hand. The vessel should be stretched just to the right tension and be positioned most favourably in the 3-dimensional peritoneal room for precise cutting. Finally the vessel should be divided by a cut in a defined area by an ultrasound cutter controlled by the left hand. Overstretching the vessel will reduce the total score; if the vessel was ripped the attempt was failed. cf. Figure 3

Materials

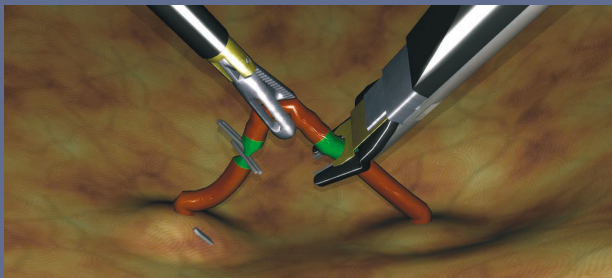
Of the basic skills tasks: Lifting and Grasping, Cutting and Clip applying were chosen; from the gynaecological module the Salpingectomy task was used. In all tasks the following were trained: Speed, accuracy, 3-dimensional perception and eye-hand coordination.

For all tasks the LapSim simulator recorded the number of attempts, time in seconds, left- and right instrument angular path in degrees, left- and right-instrument path length in meters, various error parameters like tissue damage, vessel stretch damage, misplaced or dropped clips, bleeding and blood loss, use of diathermy on non target tissue etc., and a composite score: Total Score, defined by the training course manager. The settings and requirements of the simulator are listed in the appendix page 54.

Operations in the transfer study (Paper III) were carried out in the participating department's operating theatres and were recorded from the laparoscopic camera present (Stryker®, Stryker inc., Kalamazoo, MI, USA, Storz®, Karl Storz GmbH, Tuttlingen, Germany or Olympus®, Olympus corp. Tokyo, Japan) on to a Digital Video Disc Recorder (Sony® HR-D 720, Sony Ltd. Tokyo, Japan), using the local available instruments.

Figure 4

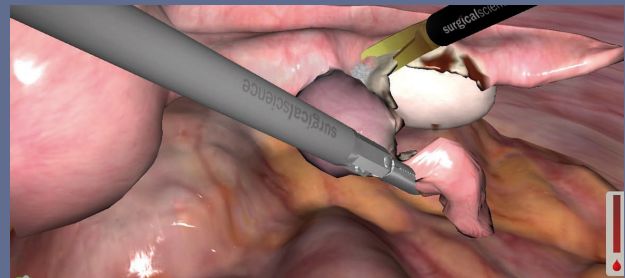
Graphic task presentation:
Clip Applying



The third basic skills task “Clip applying” trained use of clips applicator, shift of instruments and use of suction device. It also trained decision making by introducing a sudden bleeding which the surgeon had to manage. A vessel-like structure should be grasped by using left or right hand of own choice. Thereafter the surgeon had to place two clips correctly in two target areas. The surgeon should then change instrument to scissors dividing the vessel, after which a “random” bleeding starts. Instruments must be changed back to clips applicator to stop the bleeding by applying a clip proximally on each piece of the bleeding vessels. Finally the blood must be removed by switching to the suction device. cf. Figure 4

Figure 5

Graphic task presentation:
Ectopic Pregnancy



The procedural module holds all sub-elements of a complete laparoscopic procedure which were trained in the different basic skills tasks. In this task the surgeon should remove the right fallopian tube, a salpingectomy. The task entails choosing the appropriate instruments, grasping and positioning the tissue, use of diathermia only on target tissue, alternate use of scissors and diathermy. The task also trains procedural knowledge and decision making. The task setup forces the surgeon to observe certain rules in order to succeed e.g.: not to cut un-coagulated vessels or to cut the fallopian tube too far from the uterus. cf. Figure 5

Results

Paper I: Construct and discriminative validity of LapSimGyn and generating learning curves

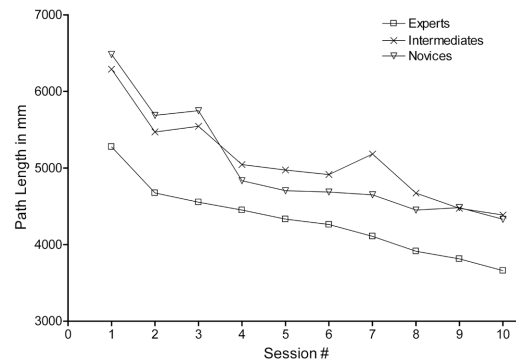
In all three Basic Skills tasks, significant differences, measured as economy of movement: Angular Path, Path Length and Total Score, were demonstrated between the expert group versus the intermediate and novice group, but not between the intermediate and novice group. Results from the intermediate experienced group were inconclusive due to a very wide range of observations, reaching from novice to expert level.

Basic skills 1

In Basic skills 1, Lifting and Grasping the Total Time differed significantly ($P < 0.016$) in the three first sessions, thereafter the Total Time tended to converge. The error parameter Maximum Damage was not significantly different between any of the groups, but Tissue Damage was significantly lower in the expert group compared to the other groups in the first three sessions ($P < 0.043$). The learning curve in this task did not tend to plateau in the valid parameters within 10 repetitions: Figure 6: Lift and Grasp: Path length in mm.

Figure 6

Lift and Grasp: Path length in mm

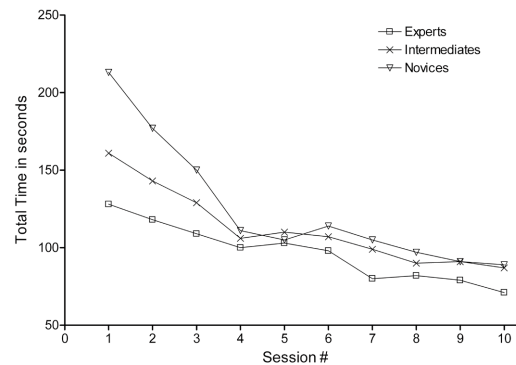


Basic skills 2

In Basic skills 2 (Cutting) significant differences were also demonstrated between the expert group and the intermediate and novice groups measured as Total Time and Total Score ($p < 0.01$) in the first 3 sessions. Experts had significantly lower scores in Tissue Damage during the first two sessions (Experts: 1.3 ± 0.47 VS Intermediates 3.4 ± 0.93 and Novices 5.3 ± 1.45 , $P < 0.04$), hereafter no significant difference was observed. There was no statistically significant difference regarding Maximum Damage, Rip Failure, Drop Failure and Maximum Stretch Damage. Figure 7: Cutting: Total time in seconds.

Figure 7

Cutting: Total time in seconds

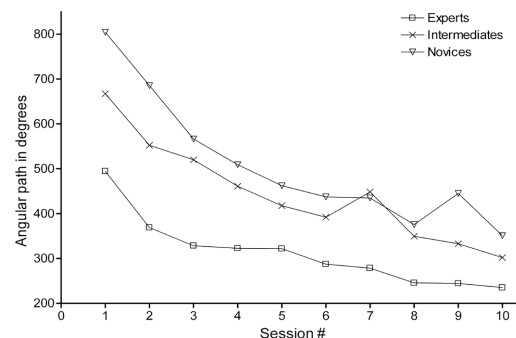


Basic skills 3

In Basic skills 3 (Clips Applying) instrument movements: Path Length and Angular Path, Total Time, Blood Loss and Total Score, the expert group performed significantly better than the less experienced groups. The rest of the error parameters showed no significant differences. Figure 8: Clips applying Angular Path.

Figure 8

Clips applying Angular Path



Results

Ectopic Pregnancy (Salpingectomy)

In the procedural module Ectopic pregnancy (Salpingectomy), significant differences between the Experts and the less experienced groups were demonstrated in Total Time, Blood Loss (Figure 9) and median Total Score (Figure 10): Instrument Path Length, and Blood Pool, Tube Cut Distance to Uterus, Ovary Damage (diathermy) and Unremoved Dissected Tissue did not differ

significantly. The learning curve is steeper and leftwards shifted in the expert group, reached their plateau after the 2nd versus the 8th iteration in the less experienced groups (Figure 11). The plateau reached by experts was more than 10% higher measured as relative score in percent compared to the less trained groups. The intermediates' median score followed the novice curves⁴⁶.

Figure 9
Box plot Ectopic Median blood loss

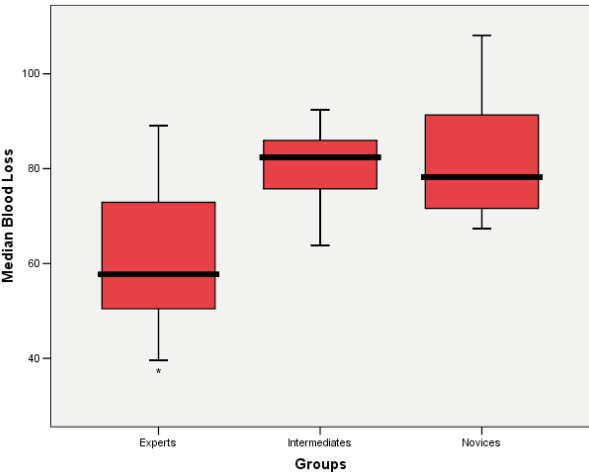
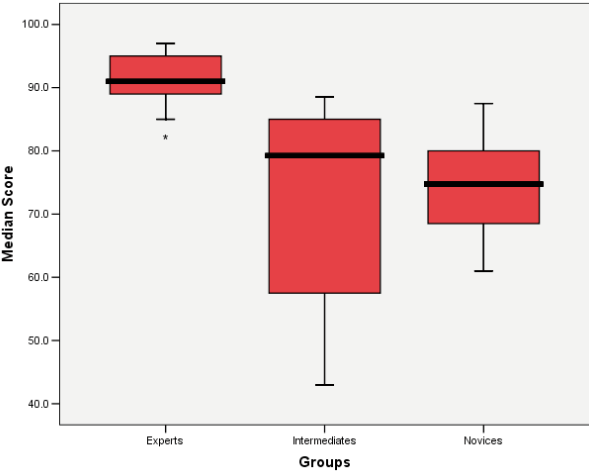
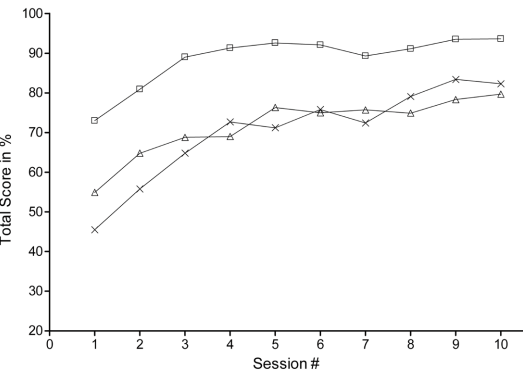


Figure 10
Box plot Ectopic Median total Score



Band: Mean, Box: 25-75 % Inter Quartile Range, Whiskers: Range.

Figure 11
Box plot Ectopic Total Score %
(Time and Motions)



Results

Paper II: Objective assessment of surgical competence in gynaecological laparoscopy: Development and validation of a procedure-specific rating scale

Construct validity

The independent and blinded evaluation of the operations resulted in a median score in the novice group of 24 (IQR 24 – 25), in the intermediate experienced group of 30 (IQR 28 – 31) and in the expert group of 40 (IQR 34 – 43). This showed that the OSA-LS was construct valid and able to discriminate between all groups, ($p < 0.03$): Table 3. The difference in overall score between the novices and the intermediates was six points (6 points round off, median 5.5 points) and between intermediates and the experts 10 points. The overall inter-rater agreement was 0.831, varying from 0.759 in the expert group to 0.905 amongst the intermediates⁴⁷. The OSA-LS assessment chart is in appendix page 54-55.

Table 3

OSA-LS: Results, Construct Validity, comparison of median, Kruskal-Wallis. Post Hoc: Mann-Whitney Bonferroni corrected

	N	Median	Interquartile Range		Range		P-value	Mann-Whitney (post Hoc)
			25% Percentile	75% Percentile	Minimum	Maximum	Kruskal-Wallis	
Novice	8	24	24	25	24	26	< 0.001	Novice vs Intermediate <0.03
Intermediate	7	30	28	31	28	33	< 0.001	Intermediate vs Experts <0.03
Expert	6	40	34	43	32	44	< 0.001	Intermediate vs Experts <0.03
All	21	31	25	34	24	44	-	

Gamma coefficient

The gamma correlation coefficient was 0.91 (95% CI 0.785 – 1.000) for all observations. The lowest correlation was found in the novice group, the highest correlation in the expert group. Even in the novice group, where the lowest gamma correlation coefficient was found, the discrepancy of the observers' ratings was randomly distributed. This emphasises that none of the observers systematically rated the performances differently from the other observer, i.e. neither more negative or nor more positive.⁴⁷

Kappa value

The Kappa value on Items level (all 21 subjects) varied from 0.510 – 0.933, items 2, 4 and 6 were the main sources of disagreement; the other items reached a higher degree of agreement. The median time used for evaluating the unedited video recordings, including filling out the score table, was 16 (range 7-35) minutes (Table 4).⁴⁷

Results

Table 4
Kappa values on single items level

Item	n	Kappa	St. error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	21	0.85	0.10	0.66	1.00
2	21	0.65	0.13	0.39	0.91
3	21	0.93	0.07	0.81	1.00
4	21	0.63	0.13	0.37	0.88
5	21	0.79	0.11	0.57	1.00
6	21	0.51	0.15	0.22	0.80
7	21	0.73	0.15	0.44	1.00
8	21	0.70	0.14	0.43	0.96
9	21	0.70	0.13	0.44	0.95
10	0	-	-	-	-

Paper III: Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial

Operative performance

Median Total Time to complete the procedure was 12 minutes (IQR 10 – 14, n=10) minutes in the simulator trained group compared to 24 minutes (20 – 29, n=11) in the control group, $p<0.001$ (Mann-Whitney) Figure 12. The median Total Score on the general and task specific rating scale reached 33 points (Inter-quartile range: 32 – 36 n=11) in the simulator trained group and 23 points (IQR 22 – 26, n=10) in the control group, $p<0.001$ (Mann-Whitney): Figure 13. A total of 21 operations were assessed.

Figure 12
Transfer of skills: Median total time

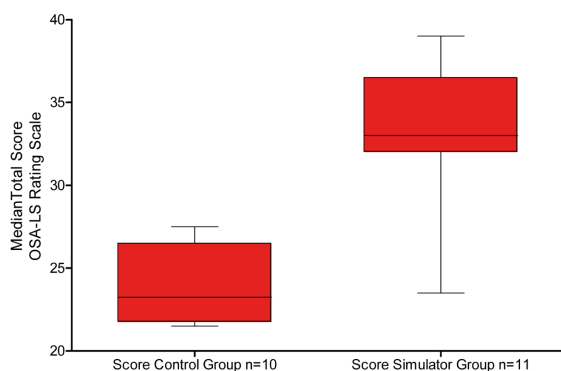
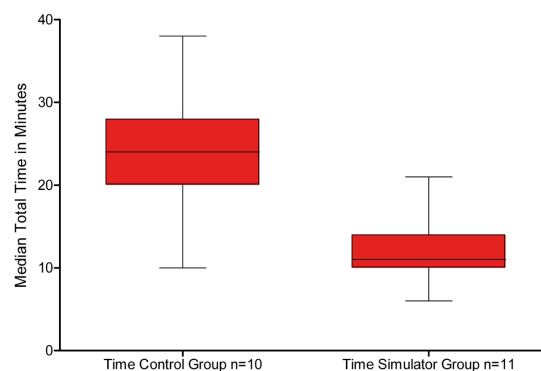


Figure 13
Transfer of skills: Median total score



Band: Median, Box: 25-75 % Inter Quartile Range, Whiskers: Range

Results

Baseline characteristics and simulator training program

The median number of simulated salpingectomies needed to reach the proficiency level in the intervention group was 28 (IQR 24 – 32, Range 16 – 39). The control group was offered simulator training after the test operation, of the 11 subjects in this group, nine volunteered for training, and used a median of 26 (IQR 23 – 32,) repetitions to reach the proficiency level, which was insignificantly different from the intervention group ($p=0.76$) The mean time used on simulator training was 7 hours 15 minutes (Range 5 hours 30 minutes – 8 hours 0 minutes) in the intervention group and 7 hours 0 minutes (Range 5 hours 15 minutes – 7 hours 45 minutes: in the control group, also a insignificant difference ($p =0.70$); Table 5. The baseline score (first attempt) was 8% (5-15%) in the intervention group and 9% (7-19%) in the control group (post operatively) which again was not significantly different ($p= 0.65$):

Table 5

Virtual reality simulator training program: Number of sessions and time used for training. Intervention group (Preoperative training) compared to postoperative training in control group. (Mann-Whitney) ns=non-significant

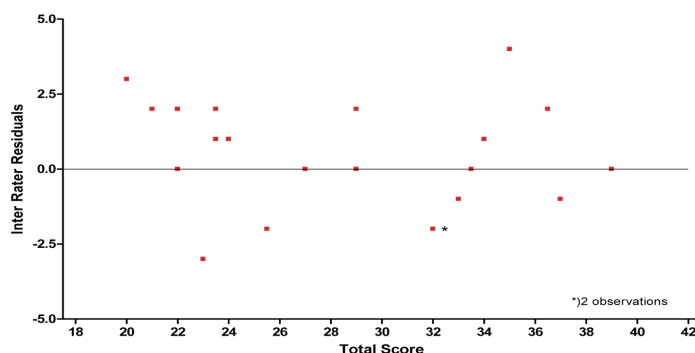
Results	Simulator: Training Sessions in numbers	Simulator: Training time in hours and minutes
Intervention group n=11 vs. Control Group (Post operative voluntary training) n=9	28 Range 16 – 39	7h 15m Range 4h 15m – 9h 30m
P-value	0.76 (ns)	0.70 (ns)

Inter-Rater Agreement & Gamma Coefficient

The time used by the assessors was the mean total operation time plus five minutes for each video, to fill in the rating chart. The inter-rater agreement calculated as the number of agreements on the assessed items divided by the total number of assessed items - was $166/210 = 0.79$ The gamma coefficient was used to investigate strength of correlations among the observers at single subject level, and reached 0.83 (95% Confidence Interval 0.69 - 0.99). The observer A minus observer B, residual score (Bland-Altman Plot) demonstrates that there was no systematic disagreement among the observers: Figure 14.

Figure 14

Bland-Altman plot Observer A minus B score



Discussion

In the present study we have been able to demonstrate that virtual reality simulation is a valid tool for assessment of laparoscopic surgical skills and that virtual reality simulation is an effective tool in the basic training of gynaecological trainees. We have also demonstrated development and validation of an objective structured rating scale for laparoscopic gynaecology.

There is an increasing interest in and need for the training and assessment process to be transparent. The use of simulation is described as an option for objective and transparent training, and as a measure of skills; and as an alternative to patient-based training and assessment³². Simulation used for assessment of technical skills have been validated in some areas of minimal invasive surgery⁴⁹⁻⁵², gynaecology⁴⁶ and in anaesthesia⁵³.

Our results support the expert opinions on simulators and provide evidence for the possible advantages of simulation in training as listed in table 2, paragraph 11, 12 and 13. The use of simulators for assessment is supported by our findings and will be discussed in the paragraph: Simulator Based Assessment of Technical Skills, page 24.

Surgical competence

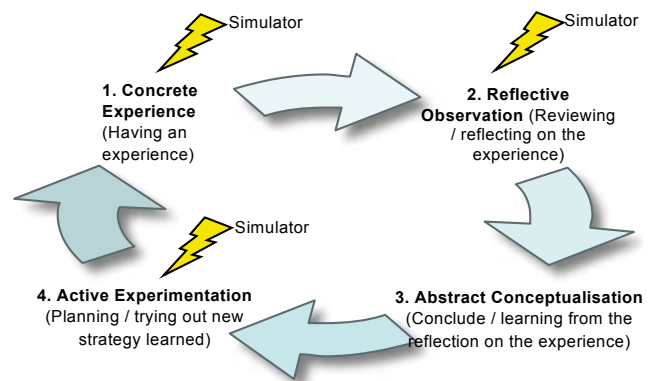
Surgical competence entails a combination of different capacities; cognitive (knowledge and decision making), behaviouristic (empathic, communicative and leadership), perceptual (visio-spatial and tactile perception) and psychomotor (technical or dexterity skills)^{4,49}. Traditional education, combined with clinical experience, is capable of providing knowledge and capacity for decision-making and communication. The perceptual and psychomotor skills are of paramount importance for surgery; however they are not optimally trained by traditional educational methods⁵⁴. We have, in the present study, focused on psychomotor technical skills, however, by using a procedural simulator, cognitive capacities such as knowledge of the surgical procedure the regional anatomy has also been trained.

Simulation and the educational theory on surgical skills learning

Currently the predominant theory on skills acquisition and improvement is based on constructivism. The basis for this theory is that a change in behaviour or/and knowledge is achieved through learning experiences. The “concrete experience” in the learning situation is believed to lead to “Reflective Observation”. At this stage the trainee reflects on her or his performance. The “Reflective observation” should then lead to an “Abstract Conceptualisation” in which the learners modify his or her strategies or behaviour in order to improve the performance. “Active Experimentation” provides new “concrete experiences” and the cycle is complete: Figure 15

Figure 15

The Kolb experiential learning cycle based on the Kurt Levin learning theory^{36,55}. Lightnings indicates possible impact points of simulator



Applied to laparoscopic skills training the Kolb learning cycle will start with the identification of the need for learning a new skill. The trainee will then handle the instruments based on procedural knowledge. Recognizing the need for improvement based on “Reflective Observation” should then lead the trainee to “Abstract Conceptualisation” in which a new improved strategy and behaviour is tested through Active Experimentation. For this cycle to function, objective feedback (for reflective observation) as well as repetitive skills rehearsals (active experimentation) are required⁵⁵.

Discussion

The virtual reality simulators should ideally provide both objective feedback and unlimited skills rehearsals. This view of the fundamental conditions for skills improvement implies that the subjects get objective assessment in order to get detailed feedback. In surgery, the Kolb theory is being supported by findings by Grantcharov et al⁵⁶ on objective assessment and constructive feedback for improvement of laparoscopic performance as well as the findings in the present investigation (Paper III)⁴⁸. Simulator training can play a role in facilitating the learning cycle.

By providing explicit experiences through simulated operations the trainee is given the opportunity to make reflective observations using the feedback from the data collected. A prerequisite is that the trainee is able to learn from his or her reflections on the experience provided by the simulator and is given the opportunity for new active experimentations, leading to new experiences. This simulator facilitated learning cycle can run infinitely only limited by the time available.

The key factor besides the possibility of unlimited rehearsals is the inherent and instant feedback. Our findings^{46;48} and those of others^{57;58} investigating learning curves support this theory. In all published studies novices improved their skills substantially and rapidly during simulator training. However 10 repetitions is not enough to reach expert level as seen both in the present study on skills transfer⁴⁸, and in our^{46;57} and other learning curve studies⁵⁹. A simulator system that does not provide feedback, typically a box-trainer-like system, may not improve learning to the same extent a virtual reality simulator does. These box-trainer requires a senior laparoscopist attending to provide the feedback⁶.

Another theory on skills learning is attributed to Donald Schön, who labels professional areas of practicing competence as “zones of mastery”⁶⁰. New challenges outside this zone trigger two kinds of reflection: Reflection in action and reflection on action. Reflection in action occurs instantly and develops continually by applying

current and past experiences and reasoning to untried situations as they occur. Reflection on action occurs afterwards. This is the process of thinking back on the unfamiliar challenge that occurred, reflect on what may have contributed to the event, and whether the behaviour was appropriate and how this situation may affect future practice⁶¹. “Reflection on action” can be a problem when learning surgical skills or other highly stressful and complex technical procedures as it can be difficult to recall the actions taken or the changed behaviour during the performance of new skills. The use of simulators, which record the actions, might provide the needed ability to review and recall the event and the skills applied.

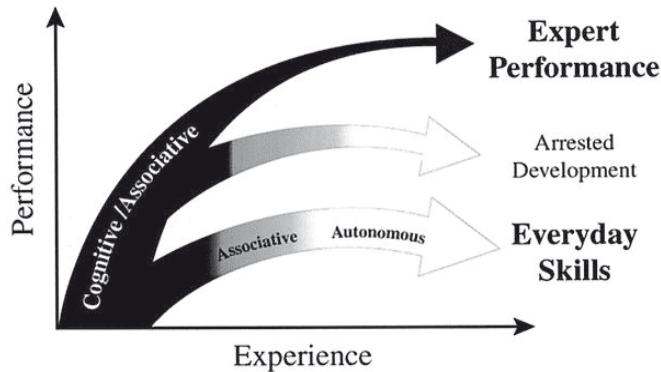
Surgical Skills Improvement

The basic learning patterns can largely be explained by the Kolb constructivist theory for learning. However the Kolb theory seems insufficient to explain why some trainees reach a higher level than others. The Kolb theory also fails to explain the situation of “arrested development”. In other words when and why does the laparoscopist’s improvement level stop and reach a plateau? Ericsson⁹ studied learning and skills acquisition, and described how to continuously improve performance of technical skills. First the subjects had to be instructed to improve some aspects of performance, secondly the subjects require immediate and detailed feedback on their performance, and thirdly the subjects had to have sufficient opportunities to improve their performance gradually through repeatedly performing the same task^{9;62}. These findings support the Kolb theory, but to reach expert level additional factors must be taken in consideration. Ericsson describes these improvements of skills in three stages: Cognitive – Associative – Autonomous (fig.16)⁶³. Normally, the goal for basic training activities of any given skill is to reach, as rapidly as possible, a satisfactory level that is stable and “autonomous.” After individuals go through the “cognitive” and “associative” phases, they can generate their performance with a minimal amount of effort⁶³ (The white arrow at the bottom of figure 16).

Discussion

Figure 16

An Illustration of the qualitative difference between the course of improvement of expert performance and of everyday activities. From K. A. Ericsson: *Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains*⁹. By permission from the author and the European Council for High Ability Studies.



Experts, however, avoid this automatism by developing increasingly complex mental representation to reach higher levels of their performance and will therefore remain within the “cognitive” and “associative” phases⁶³ (Expert Performance: top arrow in fig 16). Some (potential experts) will give up their commitment to the search for excellence and go into a “satisfactory” level that is stable and “autonomous” which results in premature automation of their performance (Arrested Development: middle arrow in fig 16)

In conclusion surgical and laparoscopic training should focus on constantly providing training of an increasingly complex level in order keep the surgeon in the cognitive/associative stage. Simulator training might

also be able to play a significant role in this part of skills improvement; however this yet has to be demonstrated. Nevertheless, this theory does not wholly explain why some of the trainees in the present transfer skills study⁴⁸ are learning faster and reach a higher performance level than others despite same goals and opportunities. This factor, the steepness of the learning curve and the level of the learning plateau is often identified as “talent” Given the limitations of this thesis we will not be able to go into a discussion of the topic of talent.

Feedback on Surgical Skills Performances

The previously presented evidence on the fundamental conditions in skills acquisition and improvement implicate that a detailed feedback is essential. The need for feedback is supported by clinical investigations such as the study “Objective assessment and constructs feedback and improvement of laparoscopic performance” by Grantcharov et al⁵⁶. Ideally, Virtual Reality simulators possess exactly the characteristics of a task objective, detailed and immediate feedback as well as repetitive task performance needed for learning and improving surgical skills. Still, in order to provide systematic feedback, objective and reliable assessment must be embedded in all training and execution of the procedure. The present transfer of skills (predictive validity) study supports this view⁴⁸. The training was exclusively based on systematic structured, objective and instant feed-back from the simulator, providing the trainee information on where, how and in which areas to improve^{46;48}.

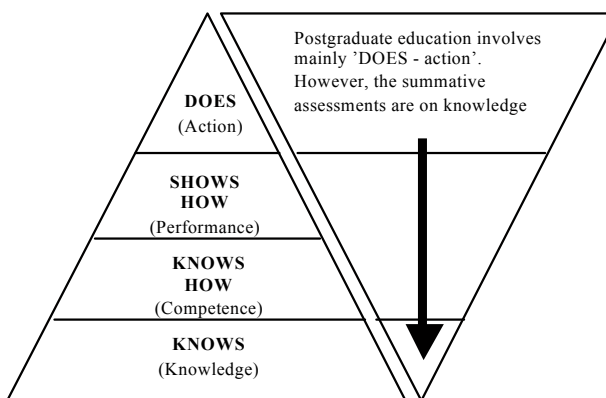
Discussion

Assessment of Technical Surgical Skills

Assessment is the key property for feedback. The choice of assessment method depends on the nature and level of what is being assessed. The level of competence and the equivalent level of assessment can be seen in the framework of Miller's triangle for clinical competence⁶⁴, with the addition of assessment levels described by Ringsted in her thesis In-training assessment⁶⁵ figure 17. Knowledge can be assessed using traditional factual knowledge tests like oral examinations, multiple choice questions and essay types of tests. Assessment of clinical competence or applied knowledge can also be tested by context-based oral examinations, multiple-choice questions or again, essay-type tests. Demonstration of skills, performance or "Shows How", can be assessed in simulated patients tests, bench station tests, or Objective Structured Clinical Examinations (OSCE)⁶⁶. Assessing the surgical skills in clinical practice, the "Does" level, can be done by observation of performance of real clinical work. The methods can be open by using direct observation, or blinded by using simulated patients, video recordings that are subsequently assessed e.g. by using checklists, Objective Structured Assessment of Technical Skills (OSATS)^{66,67}, portfolio or audits. The traditional focus in education has been on testing cognitive abilities (knowledge or Knows-How) applied as end-of training examinations, rather than assessing the actual performances (Does level), figure 17.

Figure 17

Miller's framework of competence and assessment in undergraduate education combined with Ringsted's framework for the traditional postgraduate assessment. From C. Ringsted: In-training Assessment, by permission from the author



Surgical skills are traditionally assessed in a highly subjective, unstructured and unreliable way. Competence is mainly assumed based on the number of procedures performed, the time taken to complete a procedure or in the best cases, on unstructured feedback after direct observation from a senior colleague⁵². Unlike height, weight or blood pressure, technical surgical skills cannot be determined by one simple measure. Hard endpoints like morbidity and mortality are important instruments to monitor surgical results, clinical excellence and adverse effects of specific procedures or techniques among individual surgeons and between different departments. However, these methods are unsuitable as a tool for educational feedback. First, surgical training must never jeopardize patient safety; surgical errors should be corrected immediately by the attending senior surgeon ensuring that the final result is acceptable. Secondly, surveys on surgical complications are retrospective; feedback should be instant. Thirdly, the number of surgical procedures each subject should perform to see an impact on morbidity would be too high to make patient outcome data useable for feedback or assessment of skills. Moreover surgical outcome reflects the sum of a team's work and not a single person's work.

Discussion

According to the Danish National Patient Registry for Continuous Production and Quality Control, complications demanding re-operation after laparoscopic salpingectomy in 2003 occurred in one out of 577 patients¹². In order to conduct a randomised clinical trial on the effect of skills training to improve outcome, thousands of trainees would have to be allocated to test group, which is impossible. This emphasises that adverse outcome may be used to monitor overall quality, but is inappropriate for direct educational or feedback purposes. Morbidity and mortality data are also influenced by patients' characteristics and hospital standards and environment. Morbidity and mortality data therefore do not truly reflect surgical competence at single surgeon level alone; they simply lack content validity for assessment purposes for trainees. Instead other endpoints must be engaged.

A number of different objective and systematic approaches have been presented in the literature, some based on checklists⁶⁸, some on skills performance⁶⁷, and some on error detection like time-error matrices³⁹ or on human reliability analysis⁶⁹. New technical approaches based on time and metrics have also been developed: for instance virtual reality simulators⁴⁶ and dexterity analysis systems like the Imperial Collage Surgical Assessment Device⁷⁰. The common demand for any method used for assessment, however, is that the method must be objective, reliable, valid and feasible.

Error analysis of Technical Skills

The best described error detection systems is the Time-Error matrices described by Seymour et al^{39,71} and the Observational Clinical Human Reliability Assessment system (OCHRA) by Tang et al^{69,72,73}.

Time-Error matrix

The Time-Error matrix is based on knowledge of the most frequent errors during a given surgical procedure. Seymour et al described the matrix for assessment of performance in laparoscopic cholecystectomy⁷¹. During a review of archived videotapes of laparoscopic chole-

cystectomy, Seymour et al collected potential measures of surgical performance and discussed them in an expert panel consisting of senior surgeons and a behavioural scientist⁷¹. Of all potential measures, eight events associated with the excision of the gall bladder were defined as significant deviations from optimal performance and chosen as the study measurements³⁹. The Time-Error matrix was used for observation of these predefined errors in a fixed-interval time span sampling (one – zero sampling) on a minute-to-minute basis⁷¹. In the investigations where the Time-Error Matrices have been applied the method was found to be construct-valid, and with a inter-rater agreement reaching 0.8⁴³ to 1.0³⁹. The method has also reached the same degree of inter-rater agreement in the cholecystectomy studies by Ahlberg et al.⁴³

Observational Clinical Human Reliability Assessment

In the OCHRA system errors are categorised as consequential or inconsequential according to their impact on the procedure, and subcategorised in two different types of errors; procedural and executional errors⁷². Procedural errors can be a procedural step missing or a step only partly carried out, a step unnecessarily repeated or a step out of sequence⁷². Executional errors are defined as errors in the actual performance of the surgical step. It can be a step executed with a wrongful force, speed, depth etc., or in the wrong plane or orientation, or with the wrong object. The error set is combined with checklists for all different parts of a given surgical procedure, for instance the use of electro-surgical dissection. This results in an extremely detailed assessment system, but the level of details makes the system difficult to handle. In a study by Tang et al. of 200 cholecystectomies 38,062 constituent steps were identified and 2,242 errors detected, of which 30% or nearly 700 were considered consequential⁶⁹. No doubt, this provides valuable information for research purposes of where and how errors occur during surgery, but it is obviously not a feasible system for day-to-day assessment and for feedback in general.

Discussion

Hybrid versions of the Time-Error matrix and OCHRA

The Time-Error matrix has later successfully been modified to a more detailed version⁴³, closer to the very complex OCHRA. Both systems are highly detailed, thereby very time-consuming and complicated to implement and use for the observers, which reduces their practical feasibility. Furthermore, the corrections made by the attendant surgeon in order to maintain patient safety, will (hopefully) limit the number of errors made, thereby blurring the assessment. Actually it is possible to perform a surgical procedure without errors, even though the technical performance is of low quality. On the other hand, the “near miss” situations are not necessarily recognised in a simple error detection system like the Time-Error Matrices, though, the more detailed hybrid version is more likely to detect those situations. Finally, until now there has been no work published translating the scale in to clinical experience, hence it is unclear what level a novice is expected to reach during training. Based on the reduced feasibility and the lack of translation of the scale to clinical excellence, we choose to reject these models for clinical skills assessment.

Objective Structured Assessment of Technical Skills

Based on the successful Objective Structured Clinical Examination which the Toronto-based surgical group led by Reznick developed in the late 1990's, a similar concept for the assessment of technical skills exists, called Objective Structured Assessment of Technical Skills (OSATS)^{66,67}. OSATS was originally developed for bench station tests and consisted of a task-specific checklist and a global rating scale. The global rating scale has seven items, each evaluated on a global 5-points Likert-like scale with the middle and the extreme points anchored by explicit terms⁶⁶.

It is well documented that global rating scales can be reliable, construct-valid and have high inter-rater reliability⁶⁷. Since first presented, the OSATS has been modified and tested in many different areas like open surgery⁷⁴- laparoscopy³⁸-, vascular⁷⁵- and micro-

surgery⁷⁶, urology⁷⁷, ophthalmology⁷⁸, gynaecology^{79;80} and obstetrics^{81;82}. Most tests were using OSATS as a bench station test, fewer in clinical settings. OSATS has not yet been used for laparoscopic gynaecology. In the studies referred to above, the Chronbach's coefficient alpha varied from 0.71⁶⁷ to 0.97⁸³. Chronbach's alpha expresses the reliability of the internal consistency and is discussed in detail in the discussion, page 38-39. Inter-rater reliability (IRR or inter-rater agreement: IRA) expresses the proportion of times two or more independent observers agree absolutely on their rating of a subject's performance⁸⁴. An overall agreement among two observers ≥ 0.8 based on expert opinion, is considered acceptable for a test system^{84;85}. In the studies in which inter-rater reliability is calculated, it varies from 0.70⁶⁷ to 0.97⁸⁶. Inter-rater reliability is discussed in detail in the discussion, page 38-39.

The Toronto group who created the OSATS also made a modified assessment system for the evaluation of video recordings of laparoscopic (gastrointestinal) surgery. The modified system was developed for operations on anaesthetized pigs⁸⁷. The system consists of a reduced global rating scale suitable for laparoscopic surgery, and a task-specific rating scale, called Operative Component Rating Scale (OCRS). The scale has been further modified and used for evaluation of laparoscopic cholecystectomy by Grantcharov et al. In this study as well the modified OSATS were found to be valid in discriminating different proficiency levels and with sufficiently high Inter-rater reliability^{38;88}. The solid evidence of the reliability, feasibility and construct validity of OSATS, and the modified versions for different surgical areas, have made OSATS the best documented and most widespread method for assessment of technical surgical skills. Based on these facts we chose the OSATS as the generic rating scale for our assessment system, the Objective Structured Assessment of Laparoscopic Salpingectomy; OSA-LS. The rating scale development and validation study (Paper II) confirmed that the OSA-LS rating scale was both construct- and discriminatively valid, had a high degree of observer agreement and finally that the rating scale was feasible for practical use.

Table 6
Prospective Cohort trials on construct and discriminative validity of LapSim Basic skills and the procedural modules:

Paper and year level	Modules	Sample size Subjects and level of experience	Tasks	Iterations	Outcome metrics Parameters assessed	Conclusion and comments	Evidence
Larsen et al. 2006 ⁴⁶	Basic Skills Gynaecology	N=32 Residents and specialists 11 Novice res. 0 proc. 11 Intermediate 20-60 proc. 10 Experienced > 100 proc.	1. Lifting and Grasping 2. Cutting 3. Clip apply Ectopic pregnancy	10	Time Economy of motions Blood loss Diathermy on non target tissue	Construct and Discriminatively valid on: Time, Motion, blood loss. Other parameters are not discriminative valid. Intermediate group insignificant from others, probably insufficient defined or group to inhomogeneously	2b
Aggarwal et al. 2006 ⁴⁷	Gynaecology	N=30 Residents and specialists 10 Novices res. < 10 proc. 10 Intermediate 20-60 proc. 10 Experienced > 100	Ectopic pregnancy	10	Time Economy of motions Blood loss Diathermy on non target tissue	Construct and Discriminatively valid on: Time, Motion, blood loss. Other parameters are not discriminative valid. Intermediate group insignificant from others.	2b
Ro et al. 2005 ⁴⁸	Basic skills Dissection	N=29 Residents? 13 Novices proc. 16 Intermediate < 30 proc.	1. Instrument navigation 2. Coordination 3. Lifting and Grasping 4. Cutting 5. Clip applying 6. Suturing Lap. Cholecystectomy	1	Performance score (time and errors) Economy of motions Time	No significant difference in Performance score Novices better performance in instrument navigation and suturing as well as efficiency score in dissection Inappropriate scoring system setup in this study.	3
van Dongen et al. 2007 ⁴⁹	Basic Skills	N=48 Residents? Specialists 16 Novice 0 proc. 16 Intermediate 20-60 proc. 16 Experienced > 100 proc.	1. Camera navigation 2. Instrument navigation 3. Coordination 4. Grasping 5. Lifting and Grasping 6. Cutting 7. Clip apply	?	Total score (ranking) Time Economy of motions Speed and precision Blood loss Clips placement	"...performance of the various tasks on the simulator corresponds to the respective level of endoscopic experience..." This study demonstrates construct validity for the LapSim virtual reality simulator"	≤2b
Eriksen et al. 2005 ⁵⁰	Basic Skills	N=24 Residents & specialists 14 Novice < 10 Proc. 10 Experienced > 100 Proc.	1. Camera navigation 2. Instrument navigation 3. Coordination 4. Grasping 5. Lifting and Grasping 6. Cutting 7. Clip apply	?	Total score (sum) Time Economy of motions Speed and precision Tissue damage Blood loss Clips placement Dropped clips	LapSim was able to differentiate between subjects with different laparoscopic experience. Some of the error parameters are not discriminative valid "...the system measures skills relevant for laparoscopic surgery and can be used in training programs as a valid assessment tool	≤2b
Woodrum et al. 2006 ⁵¹	Basic Skills	N=34 Med. Students, residents & specialists 9 Med students: year 1+2 20 Residents: PCY 2-5 5 Faculty: Experts	1. Instrument navigation 2. Coordination 3. Grasping 4. Lifting and Grasping 5. Cutting 6. Clip apply	10	Time Instrument pathway Instrument angular path Incomplete target areas Blood loss Time Out failure	"The LapSim has performance parameters that reliably differentiate between subjects with varying laparoscopic experience. However, some performance parameters do not differentiate between groups"	2b
Langelotz et al. 2005 ⁴⁷	Basic skills	N=115 ? & specialists 61 Intermediate < 50 proc. 54 Experienced > 50 proc.	1. Instrument navigation 2. Coordination 3. Grasping 4. Cutting & Clip apply	1	Time Instrument pathway Instrument angular path	"... it can be concluded that laparoscopic skills acquired in the operating room transfer into virtual reality. A laparoscopic simulator can serve as an instrument for the assessment of experience in Laparoscopic surgery"	≤2b

Discussion

Simulator Based Assessment of Technical Skills

Virtual Reality simulators are, for instant feedback, registering all manipulations done by the user⁸⁹. A number of different metrics are sampled such as instrument path length and angular path and time taken. Various error parameters like cauterising and cutting of non-target tissue, blood loss and un-removed dissected tissue are also collected. Data is stored in a database for documentation, log and instant feedback. The use of Virtual Reality simulators for assessment depends on the construct validity of these inherent scoring systems⁹⁰. Most commercially available Virtual Reality simulators have been investigated for construct validity in first^{91,92} second^{58,93,94}, and third generation^{95,96}, cf. Table 6: Construct validity of various Virtual Reality-simulators.

The present study (Paper I) evaluated the validity of the LapSimGyn and the LapSim Basic Skills. Both modules were found construct- and discriminatively valid on several parameters, both time and motion and certain error parameters⁴⁶. These findings have later been confirmed in additional investigations by us⁵⁷ and others^{57,58,94,96}. The concurrent results support the use of LapSim or other validated simulators for assessment of basic laparoscopic skills and performance of basic laparoscopic procedures. The high degree of discriminative validity, as demonstrated for the novices and experts, indicates that the very broad range in performance in our intermediate experienced group of junior consultants is a result of actual huge differences in performances within the group, and not a discriminative weakness of the simulator based assessment. This is supportive of the idea of using the simulator for certification in the continuous medical education.

Dexterity Analysis

for Assessment of Technical Skills

Research on expert surgical performance have revealed that economy of movements is a key difference between experts and novices^{4,98}. The fundamental principle is based on the fact that more skilful surgeons have more accurate hand and finger movements, thereby generating a more economical pattern of movements. Dexterity, in other words, is a key feature that can be measured objectively⁴. A group at Imperial College in London has developed and validated a motion tracking system (Imperial College Surgical Assessment Device: ICSAD) analysing the patterns of the surgeon's hand movements⁷⁰. The system consists of an electromagnetic field generator, and two sensors placed on the surgeon's hands. A computer collects the positional data and converts it to dexterity measures like number, speed and magnitude of hand movements, combined with time taken for the task⁹⁹. The system has been found construct- (and discriminatively) valid in both open^{70,100} and laparoscopic surgery¹⁰¹. In bench station tests Datta et al. demonstrated a strong correlation between hand motion analysis using ICSAD and OSATS global rating assessment system⁹⁸.

Discussion

Hierarchical Task Analysis of surgical procedures

In order to make an objective structured assessment of a more complex task, like a complete operative procedure, the task will have to be broken down to its least complex sub-procedures¹⁰². Hierarchical Task Analysis, or the synonymous Hierarchical Task Decomposition, HTD, is breaking down the steps of a task: each step can be decomposed into lower-level sub-steps, thus forming a hierarchy of sub-tasks. The advantage is that the decomposition into simpler sub-procedures provides knowledge of the tasks that the user wishes to perform. Hierarchical Task Analysis provides a logical rather than a psychological analysis of a task, thus it can serve as a reference against which the process functions and features can be tested³⁴. The steps of a Hierarchical Task Analysis modified from software development¹⁰³ to surgery can be as listed 1-6:

1. Decide which task to be analysed.
2. Decide upon the level of detail into which to decompose it. Making a conscious decision at this stage will ensure that all the subtask decompositions are treated consistently. It may be decided that the decomposition should continue until flows are more easily represented as a task flow diagram.
3. Break the task down into a number of subtasks. These subtasks should be specified in terms of objectives and, between them, should cover the whole area of interest.
4. List the subtasks in a layered diagram ensuring that it is complete.
5. Continue the decomposition process, ensuring that the decompositions and sequence are consistent.
6. Present the analysis to an external evaluator to check for consistency. This person should know the task at an expert level and should not have been involved in the decomposition procedure.

Modified OSATS for Laparoscopic Salpingectomy (OSA-LS)

In the present investigation (Paper II)⁴⁷ we used Hierarchical Task Analysis¹⁰² in the development of a modified global rating scale for the laparoscopic salpingectomy. In this study we first used Hierarchical Task Analysis as groundwork to define the single items of the modified OSATS. Secondly we made a comparative analysis of novice versus expert surgical performance, to define the differences between the extremes as well as the median score of the rating scale. The performance factors differentiating the expert from the novice were crucial to make the assessment system able to discriminate between different proficiency levels.

After the development of the modified OSATS rating scale for laparoscopic salpingectomy (OSA-LS) and testing the scale in a pilot study, we investigated the construct and discriminative validity in a prospective cohort study⁴⁷ (Paper II).

The OSA-LS was construct- and discriminatively valid, had a high (> 80%) Inter-rater reliability and gamma coefficient making the rating scale sufficient for general use⁴⁷. The results on validity and reliability are equal to results found in previous investigations on OSATS rating scale modified for other clinical purposes^{38;49}. We found quite a wide range in performance score among the experts. This can be explained by a wider variety of cases operated on by the expert group, i.e. a case mix situation. The novices will always have the easiest cases, and the experts the most complicated cases, such as cases with dense adhesions or significant pathological anatomy. This could influence the assessment, and might be a weakness of structured rating scales applied on heterogenic clinical work instead of homogenous bench stations⁴⁷. The rating scale was presented to an external evaluator who was not involved in the decomposition of the task to check for consistency; however a prospective validation by external experts was not performed. This impaired the external validity of the rating scale to some extent, but future use of the rating scale using external assessors might compensate for this.

Discussion

Principles for evaluating Assessment systems

Validity, Reliability, Inter-Rater Agreement, Kappa Statistic, Gamma coefficient & Kendall's Tau

Validity

Feasibility as well as validity are imperative features of any method used for objective assessment of skills. In large parts of the medical literature the term validity is interpreted as referring to the extent to which a test actually tests the competences it is designed to test¹⁰⁴. Traditionally “validity” is divided into several subtypes including *Construct validity*, *Discriminative validity*, *Concurrent validity*, *Predictive validity* and *face validity*.

Construct validity is referring to the extent to which a test actually tests what it was designed to test¹⁰⁴.

Discriminative validity is a subgroup of construct validity describing if the test is able to discriminate between proficiency levels, in other words if the sensitivity of the test is sufficient⁸⁴.

Concurrent validity refers to the extent to which the results of the test correlate with the gold standard test in the same domain⁴. E.g.: If the result in a simulator-based test of surgical skills correlates with the results of the gold standard test of the same task the tests are concurrently valid¹⁰⁴.

Predictive validity describes the extent to which a test will predict future performance, hence describing the transferability of performance from one system to another¹⁰⁴.

Face validity means the extent to which the test actually resembles what happens in real a setting⁴⁹.

The latest literature questions the traditional understanding of the division in subtypes. Cook and Beckman argue that all validity should be conceptualized under one heading: Construct validity, with all other types of validity as subgroups of this⁹⁰.

The main argument is that the assessment of psychometrics, like skills performance, has no inherent meaning, they attempt to measure an underlying construct, or an intangible collection of abstract concepts and principles. According to Cook and Beckman, this means that the result of skills assessment “has meaning only in the context of the construct they purports to assess”, hence validity is a property of inferences, not instruments, and validity must therefore be established for each intended interpretation⁹⁰. In this view the validity is not a property of a simulator or other assessment system used but of the interpretations of the measurements.

However the process of validation, both construct, discriminative and predictive, remains the same. In the assessment of a simulator system several types of validity should be investigated before considering implementing the simulator in the surgical curriculum⁴⁶. First the construct and discriminative validity of the simulator must be assessed. If valid, then the predictive validity, measured as transferability of skills obtained by training in the simulator to real operations must be assessed. If the simulator possesses predictive validity, a subject performing well in the simulator should also perform well in real operations and subjects with only mediocre performance in the simulator are also expected to have a mediocre performance in the operating room. Thus to determine predictive validity in this setting was to determine transferability of skills from simulator to operating room⁴⁸.

Reliability

Reliability is a generic term covering all aspects of reproducibility or consistency of a test¹⁰⁴. The essential notion is consistency. Consistency is the extent to which a test yields the same result when used under similar conditions or by different examiners⁸⁴. Reliability is a necessary but not sufficient component of validity. A test or assessment system that does not yield reliable scores does not permit valid interpretations⁹⁰. There are numerous ways to assess reliability; which to choose depends on the assessment instrument. In this thesis inter-rater agreement, kappa statistics and gamma coef-

Discussion

ficient were chosen to investigate the reliability of the developed rating scale.

Inter-Rater Agreement & Kappa statistics

The level of agreement between two independent observers blinded for the test subjects' training status reveals how unambiguous the test is, and thereby how reliable is the test is, if used by other independent raters. Several methods establishing the inter-rater agreement are presented in the literature⁸⁴. Cohen's Kappa value is often described as the best measure for the degree of agreement among observers. Fundamentally, Kappa calculates the degree of agreement, but it also takes into account the degree of agreement that occur by coincidence. Hence, it is argued that this statistical test is more robust⁸⁵. Such robustness is very important in a single item test, and when the outcome is binary. In a multiple-items test, or a multiple answers test like the five or more categories Likert-like scale, the overall agreement occurring by chance is negligible. Further the Likert-like scale data is ordinal: since the "distance" between the categories on the scale can differ, it is necessary to use weighted Kappa⁹⁰. In the present study (Paper II) the kappa value on the single items level had quite a broad range from poor to excellent. This was presumably due to the relatively small sample size. A simpler yet rigorous approach is to calculate the overall inter-rater agreement. This is a simple calculation done as follows:

Inter Rater Agreement = Observed events in agreement / Total number of observations⁸⁴.

In the study (Paper II) the inter-rater agreement reached 83%. That is sufficiently high for the use of the rating scale, in accordance to the statistical expert consensus, who define 80% agreement at the limit^{84;85}. A disadvantage using only Kappa statistics or inter-rater agreement is that it only provides a general value of the observer agreement. It does not make distinctions among various types and sources of disagreement. This can be investigated using gamma statistics or calculating Kendall's Tau.

Gamma Coefficient & Kendall's Tau

Data on the performance on each of the individual items included in a Likert-like rating scale is ordinal. An example is the the Likert-like scales used in many surveys: 1: Disagree strongly; 2: Disagree, 3: Neutral, 4: Agree, 5: Agree strongly. Ordinal data is categorical in which there is a logical order of the categories and the data is analysed using bivariate statistics⁸⁵. The gamma coefficient is a non-parametric rank correlation investigating agreement of ordinal categorical data which can be used to investigate strength of correlations or agreement among the observers at single subject level. Another method for investigating agreement of ordinal categorical data is Kendall's Tau rank correlation. Kendall's Tau provides also a distribution-free test of independence and a measure of the strength of dependence between two ordinal variables. Kendall Tau represents the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables⁸⁵. Gamma statistics are preferable to Spearman Rho or Kendall Tau when the data contains many tied observations, which it does in multiple assessor ratings. The gamma coefficient is calculated as the difference between the probabilities that the rank ordering of the two variables agree, minus the probability that they disagree. By applying this test we can estimate whether a possible disagreement among observers is systematic, occurs randomly or is found only in certain individuals or groups of individuals. Values of the gamma coefficient range from -1, perfect negative association to +1, perfect positive agreement. 0 indicates absence of association¹⁰⁵. In the present studies the gamma coefficient reached 0.91 in the validation of the rating scale (Paper II)⁴⁷ and 0.83 when the rating scale was applied in the transfer study (Paper III)⁴⁸. This is an acceptable gamma coefficient, which expert consensus says should be at least 0.8⁸⁵. The disagreement was randomly distributed around the line of identification indicating that there was no systematic disagreement among the observers⁴⁷. This was also confirmed in a Bland-Altman plot in the transfer study (Paper III), underlining the absence of systematic disagreement⁴⁸.

Discussion

Predictive validity

Transfer of skills

The LapSim system, the general skills module, the dissection module as well as the gynaecological procedural module “Salpingectomy”, have now been demonstrated construct and discriminative valid in several independent studies^{46;57;58;94;95;97;106}. However, the predictive validity, or the transfer of skills to operations was less well documented.

The main reasons for implementing simulators in laparoscopy training should be to train the various parts of a laparoscopic task and thereby shorten the learning curve. This should make training more efficient, less stressful, more cost effective, and in addition improve patient safety. This requires that the skills obtained by simulator training can be transferred to real life operations. Several studies have investigated this. All studies up to the present all have been carried out on laparoscopic cholecystectomy, and only the studies of Grantcharov et al³⁸, Seymour et al³⁹ and Ahlberg et al⁴³ have tested the effect of simulator training in human operations. In all other studies surrogate outcomes were used, like transfer to other simulated surgical environments, cadavers or animal models. Papers on randomised controlled trials on transferability from laparoscopic simulators to real operations and the level of evidence are listed in Table 7 a & b.

The central problem using surrogate outcomes is that it is unknown whether there is a correlation between the models and the real operations. Especially testing the training effect of one simulator in another simulator does not bring us further in answering the question of transferability to humans. Animal models resemble human operations more closely; however animal anatomy is different from human, and the stress level when performing surgery on an animal is probably much lower than in human operations. The question of transferability and impact of simulator training on the trainee's skills must therefore be addressed by assessing the actual surgical performance in real human operations.

In the present study (Paper III) we conducted a prospective randomised trial. The training of novices was carried out in the previously validated simulator LapSimGyn (Paper I), and the training was repeated until the novices reached the predefined expert level, defined in the construct validity study (Paper I). The outcome, operative performance in a human operation, was assessed using a previously validated rating scale (Paper II) The minimal clinical relevant difference in outcome and the standard deviation of the measurements from the validation of the rating scale (paper II) was used to calculate sample size. The assessment of surgical performance in the randomised trial was also done by independent observers blinded to subject and intervention group status.

The effect of the training was remarkable. The improvement in technical performance went from true novice level (24 points at the rating scale) to the performance level of intermediate experienced laparoscopists (33 points) with a simultaneous reduction in operation time of 50%. Other human studies (laparoscopic cholecystectomy) have also shown the positive effect of simulator based training^{39;43;56}. Direct comparison between the studies can be difficult due to differences in training principles, outcome measures and the interpretation of the result in terms of clinical experience.

In the study by Seymour et al. the training was criterion based, but the outcome was assessed as errors made using a Time-Error Matrix. Seymour et al. found that errors were six times less likely to occur in the simulator trained group and that gallbladder dissection was 29% faster. In contrast the non-simulator trained residents were nine times more likely to transiently fail to make progress³⁹. These are important findings but the simulator used did not train procedures, they only trained basic skills, and the assessment system used has only been applied on novices. The reference level of performances for both intermediate and experts has not been defined for the Time-Error Matrix used, making the results difficult to relate to clinical experience.

Discussion

Table 7a
Randomised controlled trials of skills transfer from virtual reality simulator to operation room, human models.

Paper and year	Power Calculation	Subjects	Simulator	Training principle	Control group	Measure (type)	Parameters assessed	Findings	Level of evidence
Larsen et al. 2008 ⁴⁶	Yes	n=24 1 st or 2 nd year Obs/gyn residents	LapSimGyn: Basic skills and Ectopic pregnancy	Proficiency based (expert criterion level)	Traditional clinical education (assistance at OR)	Human: Salpingectomy	5 general, 5 task specific items global rating: Objective Structured Assessment of laparoscopic Salpingectomy (OSA-LS) Time	"Skills in laparoscopic surgery can be clinically relevant increased by simulator training. The performance level of novices is increased to the level of intermediate experienced laparoscopists. Novices' operation time can be reduced by 50 % by simulator training".	1b
Ahlberg et al. 2007 ⁴³	Yes	n=13 surgical residents PGY1-2	LapSim Dissection	Proficiency based (expert criterion level)	Traditional clinical education	Human: Lap Chole First, fifth and tenths operation	Time Error matrices 1. Exposure errors 2. Clipping and tissue division errors 3. Dissection errors	"...Skills acquired in the LapSim improve the initial learning curve in LC, and the system is clinically validated" for this purpose. Operating time was reduced 50% in VR-trained group".	1b
Grantcharov et al. 2004 ³⁸	No	n=20 Junior surgeons	MIST-VR	Fixed number 10 repetitions of 6 tasks	No training	Human: Lap Chole	OSATS (Global rating scale) 1. Errors 2. Economy of movements 3. Time	"Surgeons who received VR training performed laparoscopic cholecystectomy significantly faster than those in the control group...and had...greater improvement in their economy of movements and error score".	2a
Seymour et al. 2002 ³⁹	No	n=16 Junior surgeons	MIST-VR	Proficiency based (expert criterion level)	Traditional clinical education	Human: Lap Chole	1. Operative errors 2. Time	"The duration of the dissection for the VR-trained group was 29% less than in the standard trained group...errors were five times more likely to occur in the standard trained group than in the VR-trained group".	2a
Hamilton et al. 2002 ⁴⁰	No	n=25 1 st and 2 nd year surgical residents	MIST-VR and Box-trainer	Fixed time 5 hours	10 VR training 9 Box training	Human: Lap Chole	OSATS (Global rating scale)	"Operative performances improved only in the VR-trained group. Psychomotor skills improve after training on both VR-trainer and Box-trainer, and skills may be transferable".	2a

Table 7b
Randomised controlled trials of skills transfer from virtual reality simulator to operation room, animal or bench models.

Paper and year	Power	Subjects Calculation	Simulator	Training	Control group principle	Measure	Parameters (Type)	Findings assessed	Level of evidence
Canal et al. 2007 ¹²	No	n=20 3 rd year medical students	Endo tower	Proficiency based (expert criterion level)	Didactic and instrument handling	Porcine: Navigation 1. Abdominal 2. Gallbladder 3. Bowel Loop	Navigation: 1. Abdominal 2. Gallbladder 3. Bowel Loop	"VR-trained subjects had better object visualization and scope orientations scores than controls... VR-trained subjects showed significant improvement in horizon scores... total error score was significantly decreased in the VR-trained group at posttest".	2a
Youngblood et al. 2005 ¹¹	No	n=46 medical students	LapSim or Box-trainer	Fixed time 3 hours (4 times 45 minutes)	No training	Porcine: Various simulated procedures	1. Time 2. Accuracy 3. Over all Global rating	"The VR-Trained group performed significantly better than the control group".	2a
Ahlberg et al. 2002 ¹⁴	No	n=29 4 th year medical students	MIST-VR	Fixed time 3 hours	Traditional clinical education	Porcine: simulated appendectomy	1. Total score 2. Grasping bowel 3. Coagulation of vessels 4. Loop ligation 5. Cutting ligature 6. Dividing bowel	"There was no significant difference between the groups, either for the total score or any part of the procedure".	2a
Hylander et al. 2002 ¹⁵	No	n=24 medical students	LapSim Basic skills	Fixed time 10 Hours (5 times 2 hours)	No training	Porcine: Various simulated procedures	1. Camera navigation 2. Instrument navigation 3. Combination 4. Extended combination 5. Time 6. Camera navigation 7. Instrument navigation 8. Combination 9. Extended combination 10. Pooled time	"The participants randomized to train with LapSim showed significantly better results for all tasks in both parts of the study than the untrained participants, according to the expert evaluator. Time consumption was accordingly lower in the training group than in the control group".	2a
Jordan et al. 2001 ^{13a}	No	n=32 (8 in VR) medical students	MIST-VR	Fixed number Six tasks	8 VR training 16 Box training 8 No training	Bench: 2 minutes paper cutting task	Accuracy: Number of correct / incorrect incisions	"Subjects who trained on MIST VR made significantly more correct incisions and fewer incorrect incisions".	2b

Discussion

In the study by Grantcharov et al. a modified OSATS was used for the assessment but in this study the training was based on fixed numbers of rehearsals, the trainees might have reached a higher performance level through further training. This could lead to an underestimation of the effect of simulator training. The simulator used was the same as in the Seymour group's study the Minimally Invasive Surgical Trainer (MIST-VR), providing abstract models for basic skills training. Nonetheless, Grantcharov et al. found a positive outcome of simulator training similar to the present study: i.e. a reduced number of errors, improved economy of movement and reduced operating time³⁸.

In the present study we found an even greater impact of simulator training, which is probably due to criterion-based training. In our study trainees had to complete on average 28 repetitions (range 16-39) to reach the defined level⁴⁸, in the study by Grantcharov et al the participants only performed ten repetitions, leaving potential room for improvement. The investigations are mutually supportive and the difference in magnitude of the training effect can be explained by the differences in simulator type and the number of repetitions performed prior to the operation.

Ideally in certification matters it is extremely important that a test is valid. Refusing a trainee with sufficient skills to operate will be inconvenient; to let a trainee surgeon with inadequate skills operate would be unacceptable. The high degree of construct validity and inter-rater agreement makes it possible that both the LapSim Virtual reality simulator and the OSA-LS rating scale can serve as a basis in high stakes situations like certification and re-certification. However, this study was not designed to define the lowest acceptable level: this will still have to be based on expert consensus according to national surgical curricula. If the assessment methods presented in this thesis are to be used for certification, further investigations to define desired cut-off performance level, must be carried out. Further, the principles from the development and validation of the OSA-LS rating scale should be applied on other frequent surgical procedures; open as well as laparoscopic, in obstetrics and gynaecology.

Another ongoing discussion among professionals and health authorities is whether assessment systems in surgery serve as a method for recruitment and career guidance. At the moment, with the simulators available, this seems to be quite difficult, based on the data from the present investigation (Paper I), and additional studies on simulator-based assessment in gynaecology⁵⁷.

Certification

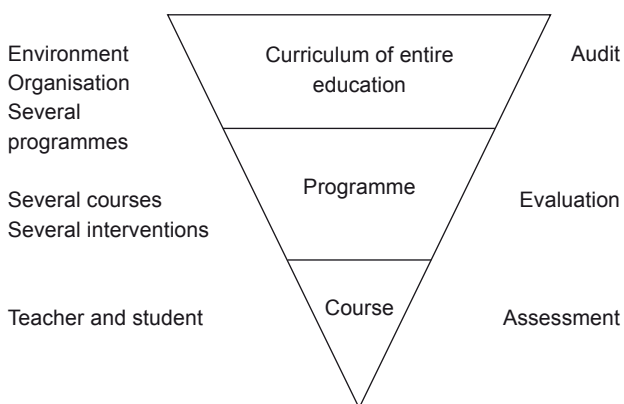
Using the OSA-LS rating Scale (Paper II), it would also be extremely difficult to distinguish talents from non-talents. The range among the novices is so narrow in this investigation that it would be impossible to distinguish unless more video recordings per individual were evaluated. However, it is yet unknown how many video recordings would be needed to obtain valid information⁴⁷. This observation is consistent with findings in the original paper where it was stated that OSATS were not designed as a predictor of surgical skills for residents before entering specialist training⁶⁷. Also the fact that all subjects in the intervention group in the transfer study (Paper III) reached the criterion level supports the view that most trainees will be able to pass, some just need more training cycles.

The great difference in the need for rehearsal emphasises how important it is to change educational strategy from fixed numbers or fixed time to criterion-based training. This is further supported by findings in related work on criterion-based training^{43;107}.

Perspectives and Future Studies

Medical education is complex, complicated, and multifactorial. Creating new educational approaches is built up by several levels. New courses will have to be evaluated and integrated into a training programme which again has to be implemented in the overall curriculum. This can be illustrated as in Ringsted: Figure 18: The relationship between training courses programmes and curricula.

Figure 18
The relationship between training courses programmes and curriculum. The present study describes the validation of the course, placed in the tip of the triangle. By permission from C. Ringsted⁶⁵.



In the present PhD project we have developed and validated a course for basic training of laparoscopic gynaecology, the tip of the triangle in figure 3. Future gynaecologists and their patients will only benefit from this laparoscopic course if it is integrated in the training programme for gynaecological trainees. From this perspective, the most important future studies are therefore on the integration of the simulator course in the training programme, and a validation of the effect of this programme. Studies using the validated rating scale to assess and give trainees feedback on their operative performance will also be important. Finally using the principles discussed in this thesis could also be used in other areas of gynaecological or other intervention spe-

cialities. Most surgical procedures still lack gold standard procedures and rating scales for structured objective feedback. There is a great potential in developing and validating training courses and assessment systems for these procedures.

The relation between simulator training and real operating practice will also need to be addressed. When to start and finish simulator training is dependent on several factors, however the retention of skills is crucial. In a follow-up study on the initial cohort of trainees included we re-rested the group twice, after 6 months break and after a further 12 months break. During the test period the participants did not perform laparoscopy neither in simulation nor in real life. It seemed that the obtained skills were kept through out the 6 months but lost some were between 6 and 12 months, without practicing (Data yet unpublished). This underlies that simulator training should be placed carefully in the surgical curriculum, and that further research on this aspect is needed.

Another issue is the number of specific procedures the trainee will have to master. At present the simulator systems automatically set limits to that, there are only a few procedures to train on. In the future, however, we must expect that number to increase and research in the optimum combination of different procedures or groups of procedures a trainee must master will have to be investigated.

In conclusion, virtual reality simulator training will be a fundamental part of future skills training. The type, amount, placement in the curriculum and implementation in the training programs is not yet clear. However, based on data from the present PhD study, from 2008, 1st year doctors in specialist training in the region of Zeeland, will have mandatory simulator-based training before performing laparoscopically on patients, and the implementation will be protocolised in order to evaluate the effect in the curriculum.

Acknowledgements

The present PhD thesis was conducted during my employment as a PhD student at the Juliane Marie Centre (JMC), for children women and reproduction, at, Copenhagen University hospital, Rigshospitalet, in the period February 2005 to 2008. I would like to express my unreserved gratitude to all co-workers, supervisors, colleagues, family and friends. In particular I would like to mention the following:

Torben V. Schroeder MD, DMSc, Professor, Department of Vascular Surgery, The Abdominal Centre, Copenhagen University Hospital Rigshospitalet, has been primary supervisor. Prof. Schroeder has done an invaluable job supporting the project, providing strategic advice as well as detailed feedback and constructive criticism of all aspects of the project and papers.

Jette Led Soerensen MD, Associated Professor senior consultant MEd, The Juliane Marie Centre (JMC), Copenhagen University Hospital Rigshospitalet, Denmark has been an inexhaustible source of ideas, inspiration, feedback, discussions and inspiration, a true driving force of research in medical education and a dedicated and sensible supervisor.

Teodor P. Grantcharov MD, PhD and Associated Professor at the Department of Surgery, University Of Toronto, ON, Canada, whose pioneering work in the field of VR-Simulation has been of invaluable inspiration. Grantcharov has not only inspired, and served as co-worker and a supervisor, but has also acted as a facilitator for collaboration with colleagues in Holland, United Kingdom, USA and Canada.

Christian Ottosen MD, Senior Consultant and **Lars Schouenborg** MD, Senior Consultant, Department of Gynaecology, JMC, Copenhagen University Hospital Rigshospitalet, Denmark for an excellent job as co-workers defining the operative procedure, conducting the Hierarchical Task Analysis and creating the OSA-LS assessment system as well as the arduous job of assessing all the video recorded operations.

Tor r Dalsgaard, PhD, Consultant, The Juliane Marie Centre (JMC), Copenhagen University Hospital Rigshospitalet, Denmark has been of great support discussing approaches to assessment of surgical competences, course development and as an observer in the operating facilities, during the data collection phase.

J rn Wetterslev PhD, Chief Physician, Copenhagen Trial Unit, Copenhagen University Hospital Rigshospitalet, Denmark has provided outstanding help to improve the protocol and managed the concealed randomisation procedure for the randomised controlled trial.

Henrik Kehlet MD, DMSc, Professor, Department of Surgical Pathophysiology, The Juliane Marie Centre (JMC), Copenhagen University Hospital Rigshospitalet, Denmark, for constructive criticism of the research protocols.

J rgen Viby Mogensen, MD, DMSc, Professor, and **Lars S. Rasmussen**, DMSc Associate professor, Senior Consultant, Department of Anaesthesiology, The Centre of Head and Orthopaedics, Copenhagen University Hospital Rigshospitalet, Denmark and their group of PhD students for inspiration and constructive criticism of the project.

Charlotte V. Ringsted, MD, PhD, MHPE, Professor, and Director of the Centre for Clinical Education, Copenhagen University Hospital Rigshospitalet, Denmark. for inspiring courses, constructive discussions and generously sharing of her knowledge on assessment topics.

Acknowledgements

Svend Kreiner, Associate professor at the Department of Biostatistics, Faculty of Health, University of Copenhagen, Denmark, for statistical discussions and support.

Susanne Tolstrup, MA, Royal Dutch Shell, Denmark
Miriam Kristensen, Journalist, Aarhus University, Denmark, for proof reading.

Anne Marie Larsen MA, LL. M., Managing Director, Hérault, France, and **Andrew Duncan Jones** BA, Hérault, France, for second proofreading.

Bent S. Ottesen MD, DMSc, Professor, Managing Director at JMC, Copenhagen University Hospital Rigshospitalet, Denmark The ideal throughout my whole professional career; for inspiration, management and strategic guidance during the process, as well as funding. The project would have been impossible to carry out without his visionary leadership.

I would also like to thank my fellow PhD students and the staff at the gynaecological departments in the region of Zeeland, Denmark, for their enthusiastic collaboration on the project. Finally I would like to thank my family for their support and tolerance, for the many hours I have been preoccupied by the project.

Reference list

- (1) Van Der Vleuten C.P.M., Dolmans D.H.J.M., Scherpbier A.J.J.A. The need for evidence in education. *Med Teach* 2000; 22(3):246-250.
- (2) Kneebone RL, Nestel D, Chrzanowska J, Barnet AE, Darzi A. Innovative training for new surgical roles--the place of evaluation. *Med Educ* 2006; 40(10):987-994.
- (3) A prospective analysis of 1518 laparoscopic cholecystectomies. The Southern Surgeons Club. *N Engl J Med* 1991; 324(16):1073-1078.
- (4) Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003; 327(7422):1032-1037.
- (5) Pickersgill T. The European working time directive for doctors in training. *BMJ* 2001; 323(7324):1266.
- (6) Issenberg SB, Scalese RJ. Simulation in health care education. *Perspect Biol Med* 2008; 51(1):31-46.
- (7) Lurie SJ. Raising the passing grade for studies of medical education. *JAMA* 2003; 290(9):1210-1212.
- (8) Halstad W S. The Training of the Surgeon. *Bull Johns Hopkins Hosp* 1904; 15:267-276.
- (9) Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med* 2004; 79(10 Suppl):70-81.
- (10) Reich H. Laparoscopic oophorectomy and salpingo-oophorectomy in the treatment of benign tubo-ovarian disease. *Int J Fertil* 1987; 32(3):233-236.
- (11) Nezhat F, Nezhat C, Gordon S, Wilkins E. Laparoscopic versus abdominal hysterectomy. *J Reprod Med* 1992; 37(3):247-250.
- (12) Lidegaard O, Hammerum MS. [The National Patient Registry as a tool for continuous production and quality control]. *Ugeskr Laeger* 2002; 164(38):4420-4423.
- (13) Karvonen J, Gullichsen R, Laine S, Salminen P, Gronroos JM. Bile duct injuries during laparoscopic cholecystectomy: primary and long-term results from a single institution. *Surg Endosc* 2007; 21(7):1069-1073.
- (14) Avital S, Hermon H, Greenberg R, Karin E, Skornick Y. Learning curve in laparoscopic colorectal surgery: our first 100 patients. *Isr Med Assoc J* 2006; 8(10):683-686.
- (15) Kumar U, Gill IS. Learning curve in human laparoscopic surgery. *Curr Urol Rep* 2006; 7(2):120-124.
- (16) Eto M, Harano M, Koga H, Tanaka M, Naito S. Clinical outcomes and learning curve of a laparoscopic adrenalectomy in 103 consecutive cases at a single institute. *Int J Urol* 2006; 13(6):671-676.
- (17) Adibe OO, Nichol PF, Flake AW, Mattei P. Comparison of outcomes after laparoscopic and open pyloromyotomy at a high-volume pediatric teaching hospital. *J Pediatr Surg* 2006; 41(10):1676-1678.
- (18) Fleisch MC, Newton J, Steinmetz I, Whitehair J, Hallum A, Hatch KD. Learning and teaching advanced laparoscopic procedures: do alternating trainees impair a laparoscopic surgeon's learning curve? *J Minim Invasive Gynecol* 2007; 14(3):293-299.
- (19) Peters J D. Cutting the legal risks of laparoscopy. *OBG MANAGEMENT* 2002; 14(10):47-55.
- (20) Carlsson S, Nilsson A, Wiklund PN. Postoperative urinary continence after robot-assisted laparoscopic radical prostatectomy. *Scand J Urol Nephrol* 2006; 40(2):103-107.

Reference list

- (21) Rosser JC, Jr., Rosser LE, Savalgi RS. Objective evaluation of a laparoscopic surgical skill program for residents and senior surgeons. *Arch Surg* 1998; 133(6):657-661.
- (22) Grantcharov T P, Funch Jensen P. Learning curve patterns in technical skills acquisition in laparoscopic surgery. Can everybody learn it? Association of Surgeons of Great Britain and England Clinical Congress <http://www.asgbi.org.uk/edinburgh/pdfs/abstracts.pdf>. 2006. Ref Type: Abstract
- (23) Johannsson H, Ayida G, Sadler C. Faking it? Simulation in the training of obstetricians and gynaecologists. *Curr Opin Obstet Gynecol* 2005; 17(6):557-561.
- (24) Jordan JA, Gallagher AG, McGuigan J, McClure N. Virtual reality training leads to faster adaptation to the novel psychomotor restrictions encountered by laparoscopic surgeons. *Surg Endosc* 2001; 15(10):1080-1084.
- (25) Jordan JA, Gallagher AG, McGuigan J, McClure N. Randomly alternating image presentation during laparoscopic training leads to faster automation to the "fulcrum effect". *Endoscopy* 2000; 32(4):317-321.
- (26) Gallagher AG, Smith CD, Bowers SP, Seymour NE, Pearson A, McNatt S et al. Psychomotor skills assessment in practicing surgeons experienced in performing advanced laparoscopic procedures. *J Am Coll Surg* 2003; 197(3):479-488.
- (27) Gallagher AG, Ritter EM, Lederman AB, McClusky DA, III, Smith CD. Video-assisted surgery represents more than a loss of three-dimensional vision. *Am J Surg* 2005; 189(1):76-80.
- (28) Bradley P, Bligh J. Clinical skills centres: where are we going? *Med Educ* 2005; 39(7):649-650.
- (29) Larsen CR, Sorensen JL, Ottesen BS. [Simulation training of laparoscopic skills in gynaecology]. *Ugeskr Laeger* 2006; 168(33):2664-2668.
- (30) Issenberg SB, McGaghie WC, Hart IR, Mayer JW, Felner JM, Petrusa ER et al. Simulation technology for health care professional skills training and assessment. *JAMA* 1999; 282(9):861-866.
- (31) Issenberg SB, McGaghie WC, Petrusa ER, Lee GD, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005; 27(1):10-28.
- (32) Kneebone R. Simulation in surgical training: educational issues and practical implications. *Med Educ* 2003; 37(3):267-277.
- (33) Fincher R.E, Lewis L.A. Simulations Used to Teach Clinical Skills. In: Norman GR, Van der Vleuten CPM, Newble DI, editors. *International Handbook of Research in Medical Education*. Dordrecht/Boston/London: Kluwer Academic Publishers; 2002. 499-535.
- (34) Patrick J. *Training: Research and Practice*. London/San Diego: Academic Press; 1992.
- (35) Soerensen J L. Masters Thesis in Medical Education: Obstetrics Skills Training [Centre for Medical Education, University Of Dundee, Scotland, United Kingdom; 2008.
- (36) Carter FJ, Schijven MP, Aggarwal R, Grantcharov T, Francis NK, Hanna GB et al. Consensus guidelines for validation of virtual reality surgical simulators. *Surg Endosc* 2005; 19(12):1523-1532.
- (37) Sutherland LM, Middleton PF, Anthony A, Hamdorf J, Cregan P, Scott D et al. Surgical simulation: a systematic review. *Ann Surg* 2006; 243(3):291-300.
- (38) Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg* 2004; 91(2):146-150.

Reference list

- (39) 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):458-463.
- (40) Hamilton EC, Scott DJ, Fleming JB, Rege RV, Laycock R, Bergen PC et al. Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills. *Surg Endosc* 2002; 16(3):406-411.
- (41) Youngblood PL, Srivastava S, Curet M, Heinrichs WL, Dev P, Wren SM. Comparison of training on two laparoscopic simulators and assessment of skills transfer to surgical performance. *J Am Coll Surg* 2005; 200(4):546-551.
- (42) Ganai S, Donroe JA, St Louis MR, Lewis GM, Seymour NE. Virtual-reality training improves angled telescope skills in novice laparoscopists. *Am J Surg* 2007; 193(2):260-265.
- (43) Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA, III et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007; 193(6):797-804.
- (44) Ahlberg G, Heikkinen T, Iselius L, Leijonmarck CE, Rutqvist J, Arvidsson D. Does training in a virtual reality simulator improve surgical performance? *Surg Endosc* 2002; 16(1):126-129.
- (45) Hyltander A, Liljegren E, Rhodin PH, Lonroth H. The transfer of basic skills learned in a laparoscopic simulator to the operating room. *Surg Endosc* 2002; 16(9):1324-1328.
- (46) 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):1460-1466.
- (47) Larsen C R, Grantcharov T P, Schouenborg L, Soerensen J L, Ottosen C, Ottesen B S. Objective assessment of surgical competence in gynaecological laparoscopy: Development and validation of a procedure specific rating scale. *BJOG* 2008; 115:908-916.
- (48) Larsen C R, Sorensen J L, Grantcharov T, Ottosen C, Schouenborg L, Dalsgaard T, Schroeder T V and Ottesen B. Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial. In review; *BMJ*
- (49) Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skills. *Eur J Surg* 2002; 168(3):139-144.
- (50) Wilhelm DM, Ogan K, Roehrborn CG, Cadeddu JA, Pearle MS. Assessment of basic endoscopic performance using a virtual reality simulator. *J Am Coll Surg* 2002; 195(5):675-681.
- (51) Aggarwal R, Grantcharov TP, Darzi A. Framework for systematic training and assessment of technical skills. *J Am Coll Surg* 2007; 204(4):697-705.
- (52) Aggarwal R, Moorthy K, Darzi A. Laparoscopic skills training and assessment. *Br J Surg* 2004; 91(12):1549-1558.
- (53) Wong AK. Full scale computer simulators in anesthesia training and evaluation. *Can J Anaesth* 2004; 51(5):455-464.
- (54) Wanzel KR, Ward M, Reznick RK. Teaching the surgical craft: From selection to certification. *Curr Probl Surg* 2002; 39(6):573-659.
- (55) Egidius H. Dialogernes og reflektionernes pædagogik. In: Kristensen HJ, Laursen PF, editors. *Pædagogik i det 21. århundrede*. 2 ed. Copenhagen: Gyldendalske Boghandel, Nordisk forlag; 2000. 93-114.

Reference list

- (56) Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc* 2007; 21(12):2240-2243.
- (57) 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* 2006; 113(12):1382-1387.
- (58) Woodrum DT, Andreatta PB, Yellamanchilli RK, Feryus L, Gauger PG, Minter RM. Construct validity of the LapSim laparoscopic surgical simulator. *Am J Surg* 2006; 191(1):28-32.
- (59) 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):150-156.
- (60) Illeris Knud. Refleksion og metalæring. Læring. Roskilde: Roskilde Universitetsforlag; 2001. 40-51.
- (61) Kaufman DM. Applying educational theory in practice. *BMJ* 2003; 326(7382):213-216.
- (62) Ericsson KA, Prietula MJ, Cokely ET. The making of an expert. *Harv Bus Rev* 2007; 85(7-8):114-121.
- (63) Ericsson KA. The scientific study of expert levels of performance: general implications for optimal learning and creativity. *High Ability Studies* 1998; 9(1):75-100.
- (64) Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990; 65(9 Suppl):S63-S67.
- (65) Ringsted CV. In-training assessment; In a work-based postgraduate medical education context. [Datawyse, Universitaire Press Maastricht, The Netherlands; 2004.
- (66) Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997; 173(3):226-230.
- (67) Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997; 84(2):273-278.
- (68) Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med* 1996; 71(12):1363-1365.
- (69) Tang B, Hanna GB, Joice P, Cuschieri A. Identification and categorization of technical errors by Observational Clinical Human Reliability Assessment (OCHRA) during laparoscopic cholecystectomy. *Arch Surg* 2004; 139(11):1215-1220.
- (70) Datta V, Mackay S, Mandalia M, Darzi A. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg* 2001; 193(5):479-485.
- (71) Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Andersen DK, Satava RM. Analysis of errors in laparoscopic surgical procedures. *Surg Endosc* 2004; 18(4):592-595.
- (72) Tang B, Hanna GB, Cuschieri A. Analysis of errors enacted by surgical trainees during skills training courses. *Surgery* 2005; 138(1):14-20.
- (73) Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A. Competence assessment of laparoscopic operative and cognitive skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World J Surg* 2006; 30(4):527-534.

Reference list

- (74) Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *Am J Surg* 2006; 192(3):372-378.
- (75) Beard JD, Choksy S, Khan S. Assessment of operative competence during carotid endarterectomy. *Br J Surg* 2007; Jun;94(6):726-30.
- (76) Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS et al. The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg* 2004; 240(2):374-381.
- (77) Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol* 2002; 167(3):1243-1247.
- (78) Cremers SL, Lora AN, Ferrufino-Ponce ZK. Global Rating Assessment of Skills in Intraocular Surgery (GRASIS). *Ophthalmology* 2005; 112(10):1655-1660.
- (79) Lentz GM, Mandel LS, Goff BA. A six-year study of surgical teaching and skills evaluation for obstetric/gynecologic residents in porcine and inanimate surgical models. *Am J Obstet Gynecol* 2005; 193(6):2056-2061.
- (80) Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol* 2000; 96(1):146-150.
- (81) Siddighi S, Kleeman SD, Baggish MS, Rooney CM, Pauls RN, Karram MM. Effects of an educational workshop on performance of fourth-degree perineal laceration repair. *Obstet Gynecol* 2007; 109(2 Pt 1):289-294.
- (82) Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol* 2006; 195(2):617-621.
- (83) VanBlaricom AL, Goff BA, Chinn M, Icasiano MM, Nielsen P, Mandel L. A new curriculum for hysteroscopy training as demonstrated by an objective structured assessment of technical skills (OSATS). *Am J Obstet Gynecol* 2005; 193(5):1856-1865.
- (84) Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003; 17(10):1525-1529.
- (85) Altman D G. *Practical Statistics for Medical Research*. 1.ed ed. Chapman & Hall/CRC; 1991.
- (86) Goff B, Mandel L, Lentz G, Vanblaricom A, Oelschlager AM, Lee D et al. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol* 2005; 192(4):1331-1338.
- (87) Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc* 2004; 18(12):1800-1804.
- (88) 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):771-779.
- (89) Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. *Surg Endosc* 2002; 16(12):1746-1752.
- (90) Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; 119(2):166-16.

Reference list

- (91) Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills. *Am J Surg* 2003; 185(2):146-149.
- (92) Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A. 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* 1998; 50:124-30.:124-130.
- (93) Sherman V, Feldman LS, Stanbridge D, Kazmi R, Fried GM. Assessing the learning curve for the acquisition of laparoscopic skills on a virtual reality simulator. *Surg Endosc* 2005; 19(5):678-682.
- (94) Eriksen JR, Grantcharov T. Objective assessment of laparoscopic skills using a virtual reality stimulator. *Surg Endosc* 2005; 19(9):1216-1219.
- (95) Ro CY, Toumpoulis IK, Ashton RC, Jr., Jebara T, Schulman C, Todd GJ et al. The LapSim: a learning environment for both experts and novices. *Stud Health Technol Inform* 2005; 111:414-417.
- (96) van Dongen KW, Tournioj E, van dZ, Schijven MP, Broeders IA. Construct validity of the LapSim: can the LapSim virtual reality simulator distinguish between novices and experts? *Surg Endosc* 2007; 21(8):1413-1417.
- (97) Langelotz C, Kilian M, Paul C, Schwenk W. LapSim virtual reality laparoscopic simulator reflects clinical experience in German surgeons. *Langenbecks Arch Surg* 2005.
- (98) Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg* 2002; 184(1):70-73.
- (99) Bann SD, Khan MS, Darzi AW. Measurement of surgical dexterity using motion analysis of simple bench tasks. *World J Surg* 2003; 27(4):390-394.
- (100) Brydges R, Classen R, Larmer J, Xeroulis G, Dubrowski A. Computer-assisted assessment of one-handed knot tying skills performed within various contexts: a construct validity study. *Am J Surg* 2006; 192(1):109-113.
- (101) Moorthy, Munz, Dosis, Bello, Darzi. Motion analysis in the training and assessment of minimally invasive surgery. *Minim Invasive Ther Allied Technol* 2003; 12(3):137-142.
- (102) Kirwan B, Ainsworth LK. *A Guide to Task Analysis*. London: Taylor and Francis.; 1992.
- (103) Preece J, Rogers Y, Sharp H, Benyon D, Holland S, Carey T. *Human Computer Interaction*. Wokingham, England: Addison Wesley Publishing Company; 1994.
- (104) Wass V, Van d, V, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001; 357(9260):945-949.
- (105) Alan Agresti. *Analysis of Ordinal Categorical Data*. Wiley; 1984.
- (106) Duffy AJ, Hogle NJ, McCarthy H, Lew JI, Egan A, Christos P et al. Construct validity for the LAPSIM laparoscopic surgical simulator. *Surg Endosc* 2005; 19(3):401-405.
- (107) Seymour NE. VR to OR: A Review of the Evidence that Virtual Reality Simulation Improves Operating Room Performance. *World J Surg* 2008; 32(2):182-188.

Appendix

LapSim Settings and requirements (Paper I & III)

Result parameter	Status	Min.	Opt. min.	Opt. max.	Max.	Weight
Total time (sec)	Enabled	130	0	150	230	15
Blood Loos (ml)	Enabled	0	0	50	180	15
Bleeding (ml/s)	Visible	-	-	-	-	-
Pool of Blood (ml)	Enabled	0	0	5	5	-
Ovary Diathermy Damage (s)	Enabled	0	0	0.1	0.5	5
Tube Cut: Uterus distance (mm)	Enabled	0	0	0	3	5
Bleeding vessels cut (#)	Enabled	1	1	1	1	-
Evacuation from the body (#)	Enabled	0	0	0	1	-
Left instrument path length (m)	Enabled	30	0	0.5	1.5	15
Left instrument angular path (°)	Enabled	50	50	100	300	15
Right instrument path length (m)	Enabled	0.9	0	2.0	3.2	15
Right instrument angular path (°)	Enabled	150	180	250	350	15

Configuration

Parameter	Value
Tubal Pregnancy Option	
Random Position	No
Position	1. Proximal
Size (min-max)	1. min
Initial Bleeding options	
Random	No
bleeding Rate (min-max)	7.max
Bleeding options	
Random	No
bleeding Rate (min-max)	7. max
Timing	
Use Exercise Timeout	No
Timeout after	off
Miscellaneous	
Disable Help	No

Appendix

OSA-LS General Skills

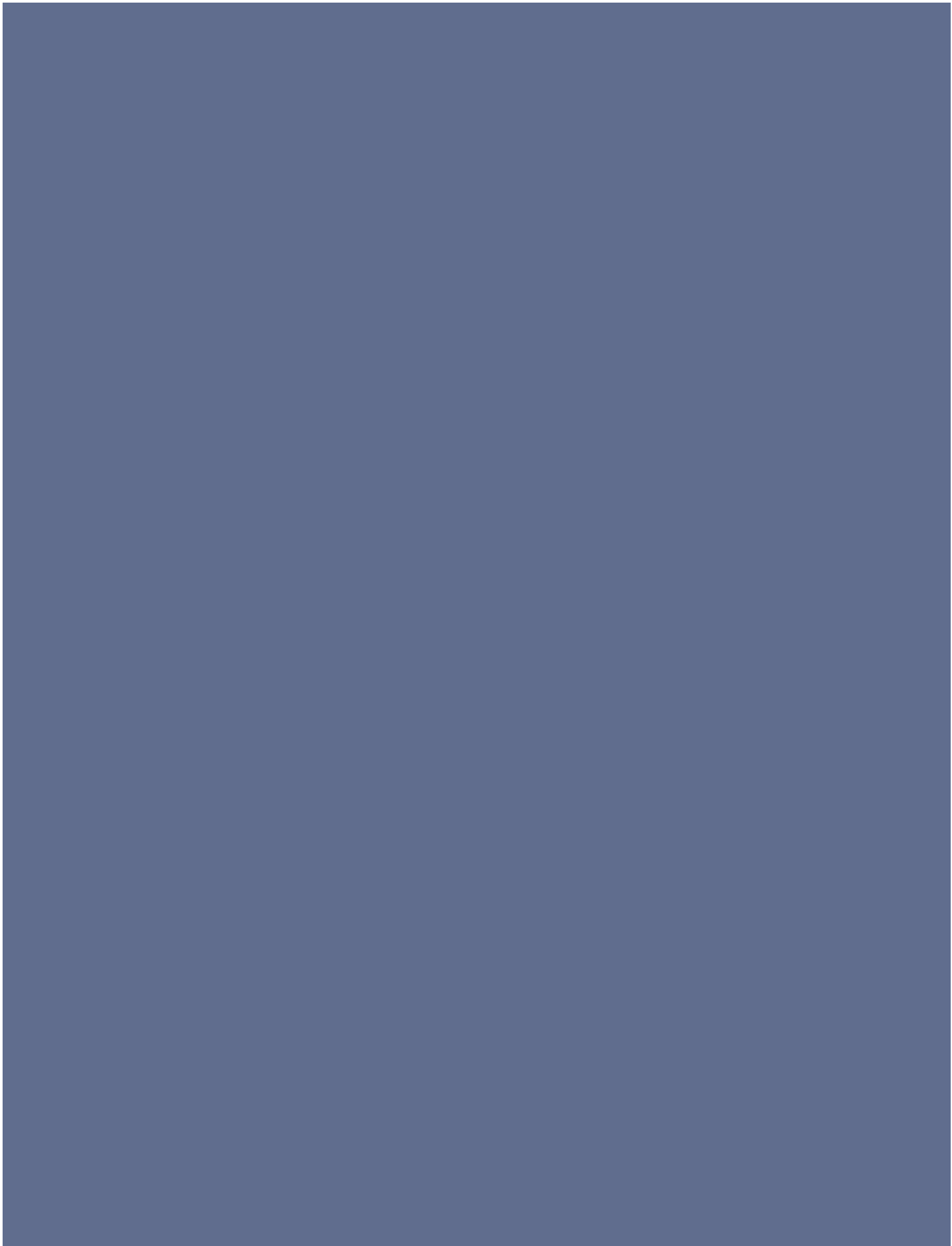
1. Economy of movements	1 Many unnecessary movements	2	3 Efficient motion but some unnecessary movements	4	5 Maximum Economy of movements
2. Confidence of movements: Instrument handling	1 Repeatedly makes tentative or awkward moves with instruments	2	3 Competent use of instruments although occasionally appeared stiff or awkward	4	5 Fluid moves with instruments and no awkwardness
3. Economy of time	1 Too long time used to perform sufficiently	2	3 Intermediate time used to perform sufficiently	4	5 Minimal time used to perform sufficiently
4. Errors: Respect for tissue	1 Frequently used unnecessary force on tissue or risk of damage by inappropriate use of instruments or instruments often out of sight	2	3 Careful handling of tissue but occasionally risk of (minimal) damage, or instruments out of sight	4	5 Consistently handled tissues appropriately with no risk of damage, instruments always in sight
5. Flow of operation / operative technique	1 Imprecise, wrong technique in approaching the operative interventions, or constant supervisor corrections	2	3 Careful technique with occasional errors, or little supervisor correction	4	5 Fluent, secure and correct technique in all stages of the operative procedure, no supervisor corrections

OSA-LS Specific Skills

6. Presentation of anatomic structures	1 Poor retraction & exposure of fallopian tube and round ligament	2	3 Satisfactory retraction & exposure of fallopian tube and round ligament	4	5 Expert retraction & exposure of fallopian tube and round ligament
7. Use of diathermy	1 Using diathermy too close to healthy ovarian or other tissue, risk of damage	2	3 Mostly safe use minimal risk of damage	4	5 Perfectly safe use of diathermy, no risk of damage
8. Dissection of fallopian tube	1 Inadequate dissection of fallopian tube. Additional damage or bleeding or part of fall. tube left <i>in situ</i>	2	3 Identified fallopian tube, Adequate dissection little damage of other structures, little bleeding	4	5 Clearly identified fallopian tube, perfectly dissected no additional damage, no bleeding. Fall. tube completely removed
9. Care for ovary, ovarian artery and pelvic wall	1 Using diathermy or cutting too close to ovarian artery high risk of bleeding or occlusion of vessel or cauterizing ovary	2	3 Mostly safe use of instruments, low risk of arterial damage, little cauterizing on ovary	4	5 Perfectly safe use instruments, no risk of cauterizing or cutting the ovary, ovarian artery or other non-target tissue
10. Extraction of fallopian tube	1 Clumsily done with major difficulty to catch the tissue, retract or get the tissue in the bag	2	3 Minor difficulty retracting or getting the tissue in the bag	4	5 Perfect retraction grasps end of structure or, easy placement of tube in bag

Paper I

Larsen C R, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, Ottesen B.
Objective assessment of gynaecologic laparoscopic skills using the LapSimGyn virtual reality simulator.
Surg Endosc. 2006 Sep;20(9):1460-6.





and Other Interventional Techniques

Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator

C. R. Larsen,¹ T. Grantcharov,² R. Aggarwal,³ A. Tully,³ J. L. Sørensen,¹ T. Dalsgaard,¹ B. Ottesen¹

¹ Department of Gynecology and Obstetrics, The Juliane Marie Centre (for children, woman, and reproduction), Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 3, DK- 2100 Ø, Denmark

² Department of Surgery, The Western Pennsylvania Hospital, Temple University Clinical Campus, 4800 Friendship Avenue, Pittsburgh, PA 15224, USA

³ Department of Biosurgery and Surgical Technology, Imperial College London, St. Mary's Hospital, Praed Street, London, W2 1 NY, UK

Received: 4 November 2005/Accepted: 2 March 2006/Online publication: 3 July 2006

Abstract

Background: Safe realistic training and unbiased quantitative assessment of technical skills are required for laparoscopy. Virtual reality (VR) simulators may be useful tools for training and assessing basic and advanced surgical skills and procedures. This study aimed to investigate the construct validity of the LapSimGyn VR simulator, and to determine the learning curves of gynecologists with different levels of experience.

Methods: For this study, 32 gynecologic trainees and consultants (juniors or seniors) were allocated into three groups: novices (0 advanced laparoscopic procedures), intermediate level (>20 and <60 procedures), and experts (>100 procedures). All performed 10 sets of simulations consisting of three basic skill tasks and an ectopic pregnancy program. The simulations were carried out on 3 days within a maximum period of 2 weeks. Assessment of skills was based on time, economy of movement, and error parameters measured by the simulator.

Results: The data showed that expert gynecologists performed significantly and consistently better than intermediate and novice gynecologists. The learning curves differed significantly between the groups, showing that experts start at a higher level and more rapidly reach the plateau of their learning curve than do intermediate and novice groups of surgeons.

Conclusion: The LapSimGyn VR simulator package demonstrates construct validity on both the basic skills module and the procedural gynecologic module for ectopic pregnancy. Learning curves can be obtained, but to reach the maximum performance for the more complex tasks, 10 repetitions do not seem sufficient at the given task level and settings. LapSimGyn also seems to be flexible and widely accepted by the users.

Key words: Assessment — Gynecology — Laparoscopy — LapSim — Simulation — Virtual reality

Over the past decade, laparoscopy has become the standard surgical approach for a majority of gynecologic procedures, such as ectopic pregnancy management, fallopian tube and ovarian operations, management of cysts, staging of gynecologic cancers, and laparoscopically assisted vaginal hysterectomy. In Denmark, 81% of all surgically treated ectopic pregnancies in 2003 were performed by laparoscopy, as compared with 64% in 1996 [11].

The advantages of minimally invasive surgery are reduced surgical trauma and thus less postoperative pain, faster recovery, shorter in-hospital stay, and better cosmetic results. However, laparoscopy demands a different set of skills than open surgery, such as the use of limited two-dimensional view over the operation field, long instruments only partly visible, impaired tactile feedback, and the fulcrum effect [1].

The traditional training and assessment of junior gynecologists are performed in an unsystematic and subjective apprenticeship model [6]. This may be adequate for open procedures, but it is not applicable for skills acquisition in minimally invasive surgery [9]. Growing public and professional concern for patient safety and increasing ethical concerns about the use of patients for training have begun to challenge this education concept. Increasing demands for operation room productivity and subspecialization together with the change toward fewer but larger hospital units concentrating on more complex surgery have further challenged the current model [14, 16].

A number of studies investigating minimally invasive general surgery have indicated that the technical skills required for laparoscopic surgery can be taught and

assessed using virtual reality (VR) simulators, provided the simulators meet fundamental requirements with respect to validity, reliability, and feasibility [7, 8].

Different educational tools such as the laparoscopic box trainer have been used to teach psychomotor basic skills. However, they lack both procedural modules and, probably most important, objective feedback and valid assessment of the trainee [13]. During the past 10 years, various graphically realistic VR laparoscopic simulators have been developed and commercially launched. They may be the tools needed for feasible, safe, and realistic training and evaluation of the younger surgeon or gynecologist [12]. There is a consensus that a mandatory feature of any training and assessment tool is construct validity, understood as the systems' ability to distinguish between different levels of performance (e.g., novice and expert) [5].

The current study aimed to investigate the potential of LapSim's basic skills and gynecologic procedural modules as valid tools for assessing psychomotor skills relevant to gynecologic surgery. Learning curves were evaluated to investigate the dose-response relationship to the training effort.

Materials and Methods

Three groups of physicians with different gynecologic laparoscopic experience were tested in the LapSimGyn v 3.0.1 Virtual Reality Laparoscopic Simulator (Surgical Science A/B, Gothenburg, Sweden). The program was executed on an IBM T42 computer in a docking station (Pentium M 1.8 GHz / 512 MB RAM; IBM Inc., Armonk, NY, USA) using a Virtual Laparoscopic Interface with a diathermy pedal (Immersion Inc., San Jose, CA, USA).

The participants were either gynecologic trainees or consultants (juniors or seniors). They were assigned to three groups of 10 subjects each: novices, with no experience in advanced laparoscopy; intermediates, defined as consultants who had conducted more than 20 and fewer than 60 advanced laparoscopic procedures during the preceding 2 years; and experts, senior consultants currently performing mainly laparoscopic surgery who had carried out more than 100 advanced operations during the preceding year. Advanced laparoscopic gynecologic operations were defined as all procedures except diagnostic laparoscopy and sterilization.

Demographics and logistics

The expert group included seven men and three women, average age of 54 years (range, 45–60 years), one left-handed, and one ambidextrous. The intermediate group consisted of five men and five women, average age of 46 years (range, 39–55 years), all right-handed. The novice group comprised three men and seven women, average age of 34 years (range 27–42 years), one left-handed, and one ambidextrous. None of the participants had previous experience with the LapSim simulator.

Introduction to the simulator

Initially, all test participants had 5 min of didactic and hands-on introduction to the LapSim system and the four different tasks, including a presentation of the parameters measured for assessment. Immediately after the introduction, the first two simulations were carried out. Each of the following two sessions were conducted as 2 × 2 simulations separated by a 30-min break. There was a 2- to 7-day break between the next two sessions. The total training period ranged from a minimum of 5 to a maximum of 15 days. Participants were excluded if the time limit of 7 days between sessions was exceeded.

Simulations were excluded and redone if technical or software problems occurred. The exclusion of results at a premature exit was automatically registered in the result database.

Of the 32 participants included in the study, 10 were novices (no dropouts), 11 were in the intermediate group, and 11 were experts. One participant from the expert group exceeded the time limit between the second and third sessions and was excluded. One participant from the intermediate group was excluded because of a broken interface handle. Consequently, each of the three groups consisted of 10 participants.

Tasks description

All the participants carried out three basic skills tasks in the same order of progressive complexity (lifting and grasping [B1], cutting [B2], and clips applying [B3]) and one procedural task (a right side salpingectomy for an ectopic pregnancy). All the participants completed the 10 sessions of all four tasks.

The basic skills

Lifting and grasping (B1) is a coordination task (involving speed and precision) in which the tested subject must lift a virtual object (a plate) with one hand and move a needle-shaped object with the other hand. There were five objects to be moved on each side. The parameters measured were Total Score (%), Total Time (s), left and right instrument Path (mm), left and right instrument Angular Path (degrees), Tissue Damage (#), and Maximum Damage (mm).

The second basic skills task (B2) was cutting, in which a small vessel was grasped by the right hand and stretched until the target area appeared. The target area then was to be cut by an ultrasonic cutter held in the left hand. Free segments were to be cut off and placed in a defined target area five times. The results were displayed as in the first basic skills task, with the additional measures Rip Failure (%), Drop Failure (%) and Maximum Stretch Damage (%).

The clips applying task (B3) required the placement of two endoclips at predefined target sites on a small vessel held by a grasper. When the clips were placed correctly, a target site at the middle of the vessel appeared, which had to be transected by a scissors. The simulator was set to start bleeding from both pieces of vessels at the site of the clips after the transection. To obtain hemostasis, another two clips had to be applied proximally from the first clips, one on each part of the vessel. To the parameters Total Time and Economy of Movement were added Incomplete Target Areas (#), Badly Placed Clips (#), Dropped Clips (#), and Blood Loss (ml).

The ectopic pregnancy procedural module

The ectopic pregnancy procedural module presented the subjects with a high-fidelity graphic image of the internal female genitals seen from a camera inserted through the umbilicus. An ectopic pregnancy was placed in the right fallopian tube, and the test participant had to perform right side salpingectomy using a grasper, bipolar diathermy, scissors, endobag, and wash/suction. The size of the fallopian tube pregnancy and the bleeding rate were set to the maximum (7 in arbitrary units), and the performance was measured as Total Score, Total Time, and Economy of Movement, as for the basic skills tasks. The additional measures were Blood Loss (ml), Blood Pool Volume (left intraperitoneal, ml), Ovary Diathermy Damage (s), Uterus Tube Cut Distance (mm), and Unremoved Dissected Tissue (#).

Simulator settings

The simulator generated the total score in percentage by adding the weighted scores of the Total Time and Economy of Movement parameters. In this trial, we chose to sum only the time and the instrument movements (left and right instrument Angular Path and Path Length), which proved to be valid for previously tested VR simulators [4]. The left and right side instrument movements were added to a sum of the total instrument movements in length and angular path for each task. Calibration of the simulator was performed

Table 1. Construct validity for basic skills 1 (lifting and grasping) and 2 (cutting)^{a,b}

Basic skills	Session	Time	Path length	Angular path	Total Score	Tissue damage	Max damage	Rip failure	Drop failure	Max stretch damage
Skill 1: lifting and grasping	1	0.016	0.021	0.005	0.011	0.016	0.299	IP	IP	IP
	2	0.013	0.019	0.015	0.019	0.012	0.168	IP	IP	IP
	10	0.136	0.004	0.002	0.002	0.231	0.984	IP	IP	IP
Skill 2: Cutting	1	0.002	0.023	0.014	0.002	0.026	0.450	0.596	0.239	0.631
	2	0.012	0.048	0.009	0.018	0.041	0.393	1.000	0.240	0.670
	10	0.164	0.022	0.200	0.032	0.438	0.642	0.596	0.902	0.449

Max, maximum; IP, invalid parameter

^a First and second simulation, with all parameters measured. The 10th session was added for comparison

^b Significant values are in bold type; *p* values are according to Kruskal–Wallis nonparametric analysis of variance

Table 2. Construct validity for basic skill 3 (clips applying)^{a,b}

Basic skill	Time	Path length	Angular path	Total score	Max stretch damage	Incomplete target areas	Badly placed clips	Dropped clips	Blood loss
Skill 3: clips applying	1	0.004	0.011	0.017	0.077	0.243	0.595	0.566	0.035
	2	0.002	0.014	0.018	0.002	0.595	0.595	0.792	0.018
	10	0.145	0.285	0.019	0.038	0.988	0.189	0.328	0.328

Max, maximum

^a First and second simulation, with all parameters measured. The 10th session was added for comparison

^b Significant values are in bold type; *p* values are according to Kruskal–Wallis nonparametric analysis of variance

after a pilot study. A complete list of requirements and settings can be obtained by contacting the corresponding author. To ensure that all subjects were given the possibility to finish each task, time out failure was disabled, and thereby not tested.

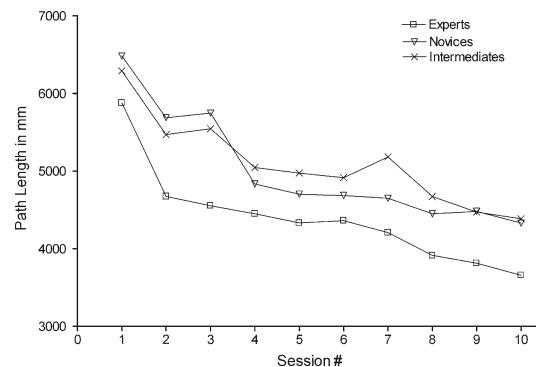
Statistical analysis

Data were analyzed using the SPSS 12.0.1 for Windows. The Kruskal–Wallis nonparametric analysis of variance was used to investigate differences among the three groups. Learning curves were generated in MS Excel 2002. Values are presented as median (range). Significant differences were investigated using the Mann–Whitney test. The *p* values were Bonferroni adjusted to reduce the risk for type 1 error. All *p* values (two-tailed) less than 0.05 were considered statistically significant.

Results

In all three Basic Skills tasks, significant differences were demonstrated between the expert group and the intermediate and novice group, but not between the intermediate and novice group, measured as economy of movement: *Angular Path*, *Path Length* and *Total Score* (Tables 1 and 2).

In the B1 task (lifting and grasping) the Total Time differed significantly ($p < 0.016$) in the first three sessions, then equalized thereafter. The error parameter Maximum Damage was not significantly different between any of the groups, but Tissue Damage was significantly lower in the expert group than in the other groups in the first three sessions ($p < 0.043$). The learning curve in this task did not tend to plateau in the valid parameters within 10 repetitions (Fig. 1: lifting and grasping Path Length in millimeters).

**Fig. 1.** The B1 task (lifting and grasping): Path Length in millimetres.

In the B2 task (cutting), significant differences also were demonstrated in Total Time and Total Score between the expert group and the intermediate and novice groups (Fig. 2: cutting Total Time). The experts had significantly lower scores in Tissue Damage during the first two sessions: experts (1.3 ± 0.47) vs intermediates (3.4 ± 0.93) and novices (5.3 ± 1.45) ($p < 0.04$). Thereafter, no significant difference was observed (Fig. 3: cutting Tissue Damage). There were no statistically significant differences in terms of Maximum Damage, Rip Failure, Drop Failure, or Maximum Stretch Damage (Table 1: construct validity, basic skills 1 and 2).

In the B3 task (clips applying), the expert group performed significantly better than the less experienced groups in instrument movements (Path Length and

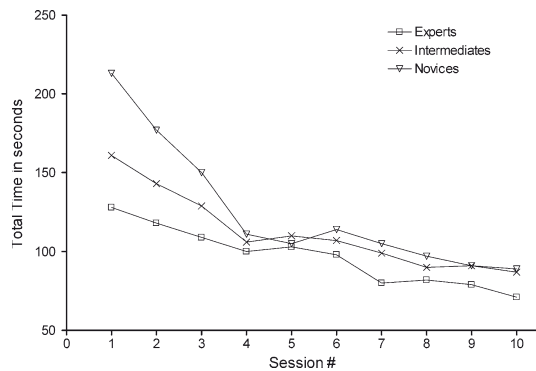


Fig. 2. The B2 task (cutting): Total Time in seconds.

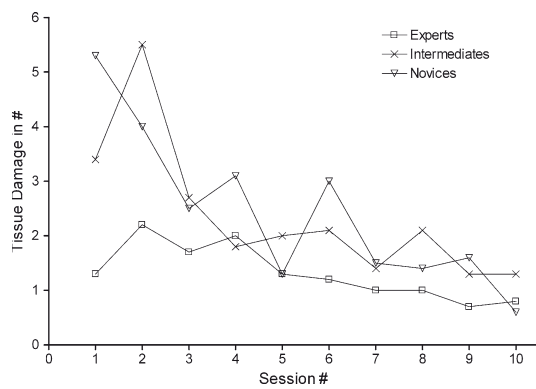


Fig. 3. The B2 task (cutting): Tissue Damage in #.

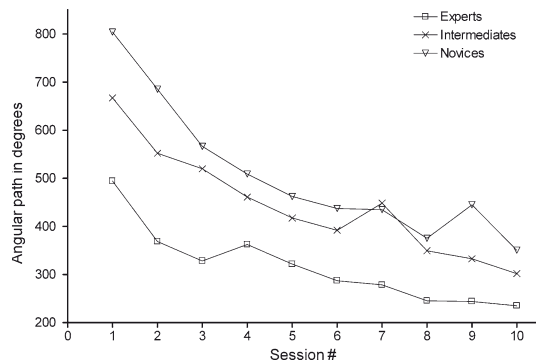


Fig. 4. The B3 task (clip applying): Angular Path in millimetres.

Angular Path), Total Time, Blood Loss, and Total Score. The remainder of the error parameters showed no significant differences (Table 2: construct validity, basic skill 3; Fig. 4: clips applying, Angular Path).

In the ectopic pregnancy task, significant differences between the experts and the less experienced groups were demonstrated in Total Time (Fig. 5), Blood Loss (Fig. 6), and Total Score (Fig. 7). The differences in instrument Path Length, Blood Pool, Uterus Tube Cut

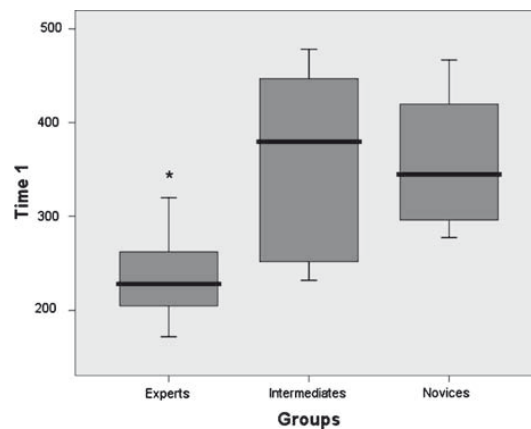


Fig. 5. The ectopic pregnancy task: Box plot of the three groups for Total Time required to complete the first simulation.

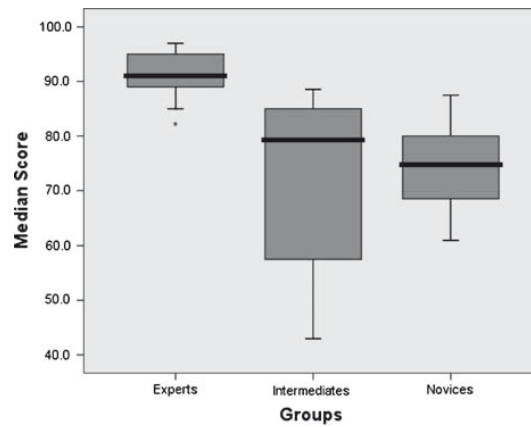


Fig. 6. The ectopic pregnancy task: Box plot of the three groups for Total Score for all sessions of the simulated salpingectomy.

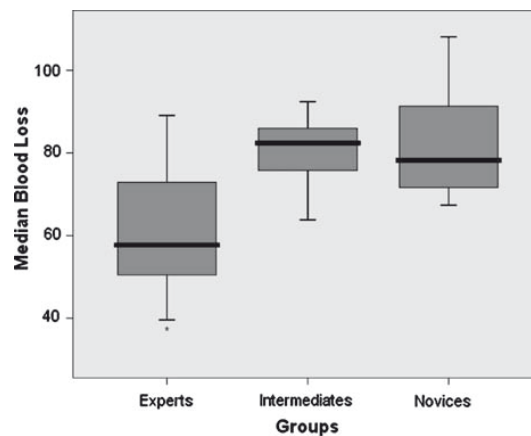
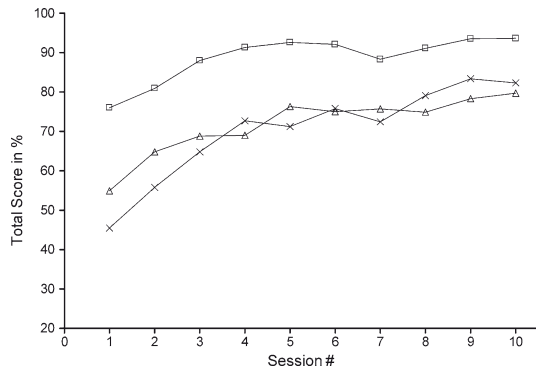


Fig. 7. The ectopic pregnancy task: Box plot of the three groups for Blood Loss for all sessions of the simulated salpingectomy.

Table 3. Construct validity for the first and second salpingectomy simulation^{a,b}

Ectopic pregnancy	Session	Time	Path length	Angular path	Total score	Blood loss	Tube cut distance	Ovary damage	Blood pool	Unremoved tissue
Salpingectomy	1	0.001	0.023	0.155	0.030	0.005	0.155	0.254	0.636	IP
	2	0.008	0.043	0.175	0.035	0.066	0.453	0.217	0.474	IP
	10	0.039	0.041	0.025	0.016	0.054	0.185	0.089	0.318	IP

IP, invalid parameter

^a All parameters were measured. The 10th session was added for comparison^b Significant values are in bold type; *p* values are according to Kruskal–Wallis nonparametric analysis of variance**Fig. 8.** The ectopic pregnancy task: Learning curve total score in %. Median values.

Distance, Ovary Diathermy Damage, and Unremoved Dissected Tissue all were insignificant (Table 3).

Learning curves

A common characteristic of all the learning curves for all the tasks and groups was that all valid parameters showed improvement in performance. Depending on the difficulty of the task, novices plateaued after the seventh or eighth repetition, except for lifting and grasping, in which there was a continuous climbing curve in Total Score and a descending curve in instrument movement (Fig. 1). In the most complex task, the ectopic pregnancy, the experts demonstrated a very short learning curve, as compared with both novices and intermediates, reaching a score of almost 90% after only two sessions and not increasing significantly thereafter. The learning curves were steeper and shifted leftward in the expert group, plateauing after the second repetition, as compared with the eighth repetition in the less experienced groups (Fig. 8). The plateau reached by the experts was more than 10% higher in terms of relative percentage score, as compared with the less trained groups.

Discussion

The current study demonstrated that the LapSim VR simulator, in both the basic skills and ectopic pregnancy

modules, was able to discriminate between expert gynecologists and others with less laparoscopic experience. Most studies of VR simulators have been conducted with surgeons. It generally is difficult to compare gynecologists with general surgeons because surgery is only a minor part of the work performed by gynecologists and obstetricians, whereas it is the primary field of a surgeon's work.

To date, only three studies investigating the construct validity of the LapSim have been published, all focused on the basic skills module, but none on the LapSimGyn procedural module. Our results for the basic skills module are consistent with those of formerly published studies [3, 10, 15]. In a Canadian trial on the basic skills module, three groups were tested: an expert group, an intermediate group (junior) consisting of surgical residents, and a novice group consisting of medical students. Consistent with our data for the same tasks (grasping, cutting, and clips applying), Sherman et al. [15] demonstrated significant differences between the expert and the two less experienced groups, except for the economy of motion score, which they did not find valid.

Despite the accordance in results, a significant advantage of the current study is that it included only physicians from the gynecologic/obstetric specialty to secure relevance for clinical practice and homogeneity for the group of novices. No studies found a significant difference in performance between the novices and the intermediately experienced. A possible explanation is that the intragroup variability among the intermediates in this study may have been too broad because of fundamental differences in the current laparoscopic workload.

To date, all the studies on the LapSim have shown that the error parameters are not valid for discriminating between different levels of experience. Possible explanations for this might be that these parameters are not valid, that the difficulty level of the tasks is too low to detect differences, and that a large number of subjects are necessary to detect statistically significant differences (type 2 error).

All three explanations actually can be true. The Unremoved Dissected Tissue parameter simply is not valid, both because it is quite simple even for novices to dissect the tissue in toto and because it is much too easy to handle the Endobag (the bag only has to touch the tissue) for removal of the dissected tissue. The requirements and settings for the Uterus Tube Cut Distance parameter were too easy in this particular setting. A

setting that met the subject with more accurate and narrow demands of precision may have made this parameter valid. The Rip Failure occurred twice as often in the intermediate and novice groups as in the expert group, and after the sixth repetition, the experts had no more episodes of Rip Failure. However, at this sample size and given the frequency of Rip Failure, no significant difference was found. This could be interpreted as a type 2 error.

Even if not valid for assessment, the nonvalid error parameters may be useful tools in training. The quality of the surgical result is not based only on time and economy, so the error parameters may help to ensure that the procedure is performed in accordance with the “gold standard procedure.” By using the error parameters, the educator can ensure that the trainees are carrying out the surgical procedures in the same way. In procedural modules, such as the ectopic pregnancy program, it is mandatory that the procedure be carried out correctly. Uterus Tube Cut Distance and Ovary Diathermy Damage are therefore important checkpoints for ensuring that a correct procedure is performed (e.g., in performing salpingectomy, the entire fallopian tube must be removed). The Uterus Tube Cut Distance is therefore a cardinal point in this module, although it is not a valid parameter in itself for discriminating different levels of experience. The Ovary Diathermy Damage parameter could be used as an extra pass/fail check so that the trainee cannot pass if diathermy is used wrongly on the ovary.

Learning curves express the relationship between an outcome variable and experience with a given task. The curve can be used to determine the learning profile and the time at which further iterations do not increase performance. It also can be used to analyze factors influencing the learning process. In the current study, inter- and intrasubject variability rapidly declined from the first session and throughout the following repetitions, showing a regimentation of the technical performance.

The plateaus varied in the different tasks. In two of the tasks (clips applying and ectopic pregnancy), we found one plateau, and in two tasks (lifting and grasping and cutting), we found no definite plateau. In the latter, we found a steep improvement of performance at the end of the trail, suggesting that maximum performance had not been reached, and that more repetitions would have been necessary to achieve a plateau in the score parameters.

Brunner et al. [2] evaluated the literature for the number of trials needed to reach the maximum plateau, investigating performing curves with 30 repetitions using Minimal Invasive Surgical Trainer (MIST). Their data demonstrate that an initial plateau is reached after 8 repetitions, but that the overall best score was reached at 21 to 29 repetitions. This suggests that more iterations than the usual 10 may be beneficial, but obviously does not show whether there is another increase in performance with more than 30 repetitions. The current study demonstrated a slightly different learning curve profile in lifting and grasping as well as cutting, with a plateau at the eighth repetition, then an increase again at the end of the trail (10th iteration), leaving an uncertainty about where to find the final plateau of performance.

The traditional surgical training is largely quantitative. It is time-based training, which means that the same amount of training is provided for a group of subjects either as a fixed number of repetitions or as a random number of repetitions during a fixed time. In proficiency or outcome-based training, individualized amounts of training (repetitions) are provided though the same level of skills. If surgical curricula are based on arbitrary numbers of repetitions, some trainees will be overtrained and waste time, whereas others would gain from further training. In the ectopic pregnancy module, although a stable plateau was reached by the less experienced, the experts still were performing significantly better, which may have made it possible for novices to improve performance further through more iterations. This emphasizes the need for a qualitative approach to skills training in which the performance of an expert, or reference group, defines the level of skills that the novices must reach before progressing to the operating theater. This approach could be the rationale for both certification of trainees and recertification of consultants.

Introducing simulator training and assessment

Despite the high cost, the potential use of the high-fidelity VR simulator as a training and assessment tool to improve clinical excellence makes it valuable for teaching hospitals. In this study, on the basis of pilot trials, we customized the settings and requirements to gain optimal knowledge on the construct validity of the simulator. The Total Score was based only on the Time and Economy of Motion parameters, which in other studies have proved to be valid at an individual level. Because most of the error parameters did not prove to be valid, this approach turned out to be correct. On the other hand, some error parameters should be included to make it impossible for trainees to pass a test if certain cardinal points are not met.

Because of the complexity and demands for accurate calibration of the LapSim system, it seems most appropriate to install and operate the simulator in central units or skills centers. It takes time for a trainee to become familiar with all the options in settings and requirements, to operate the simulator, and finally to interpret results. From this perspective, simulator training needs to be integrated into the surgical curriculum.

Conclusion and perspectives

The LapSimGyn VR simulator package demonstrates construct validity, in both the basic skills module and the procedural gynecologic model of ectopic pregnancy, even if subjects are tested consecutively in the modules. Learning curves can be obtained, but to reach maximum performance in the more complex tasks, 10 repetitions do not seem sufficient at the given task level and settings. LapSimGyn also seems to be flexible and widely accepted by users. Further studies need to investigate the predictive validity and transferability of basic and procedural skills obtained by simulator training.

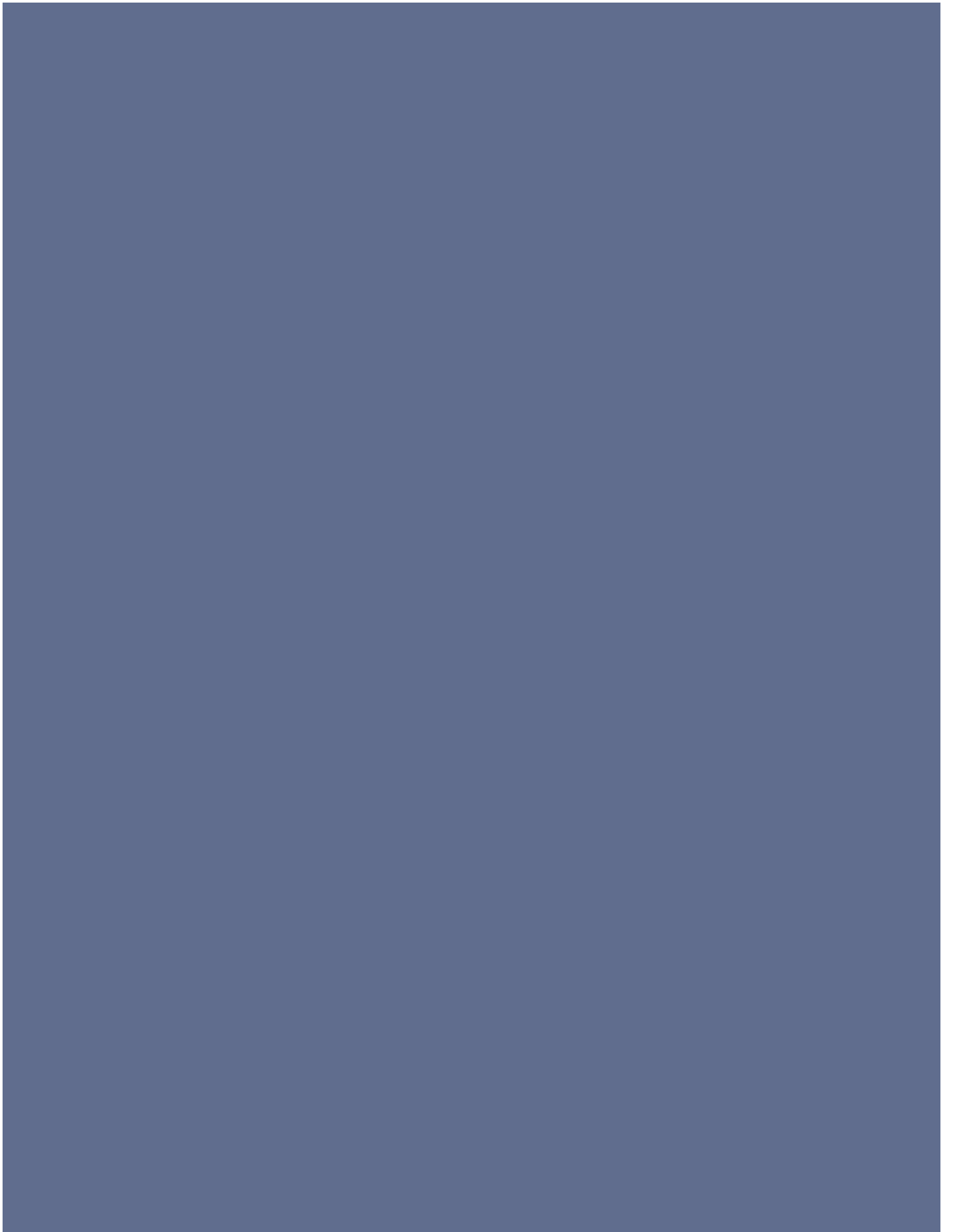
Acknowledgments. The authors thank TrygFonden and The Juliane Marie Centre for financial support, and Dr. Christian Ottosen, Dr. Lars Schouenborg, and Professor Torben V. Schroeder for inspiration and criticism. Finally, they thank the gynecologic departments at the Rigshospitalet, Frederiksberg, Hillerød, Glostrup, Hvidovre, Holbæk, and Herlev hospitals for recruiting test persons.

References

1. Aggarwal R, Moorthy K, Darzi A (2004) Laparoscopic skills training and assessment. *Br J Surg* 91: 1549–1558
2. Brunner WC, Korndorffer JR Jr, Sierra R, Massarweh NN, Dunne JB, Yau JB, Scott DJ (2004) Laparoscopic virtual reality training: are 30 repetitions enough? *J Surg Res* 122: 150–156
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, Ritter EM, Satava RM (2003) Fundamental principles of validation and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 17: 1525–1529
6. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J (2002) Assessment of technical surgical skills. *Eur J Surg* 168: 139–144
7. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J (2003) Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills. *Am J Surg* 185: 146–149
8. Grantcharov TP, Rosenberg J, Pahle E, Funch-Jensen P (2001) Virtual reality computer simulation. *Surg Endosc* 15: 242–244
9. Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS, Jarvi KA (2004) The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg* 240: 374–381
10. Hyltander A, Liljegren E, Rhodin PH, Lonroth H (2002) The transfer of basic skills learned in a laparoscopic simulator to the operating room. *Surg Endosc* 16: 1324–1328
11. Lidegaard O, Hammerum MS (2002) The National Patient Registry as a tool for continuous production and quality control. *Ugeskr Laeger* 164: 4420–4423
12. Moorthy K, Munz Y, Sarker SK, Darzi A (2003) Objective assessment of technical skills in surgery. *BMJ* 327: 1032–1037
13. Munz Y, Kumar BD, Moorthy K, Bann S, Darzi A (2004) Laparoscopic virtual reality and box trainers: is one superior to the other? *Surg Endosc* 18: 485–494
14. Reznick RK (2005) Surgical simulation: a vital part of our future. *Ann Surg* 242: 640–641
15. 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
16. Wanzel KR, Ward M, Reznick RK (2002) Teaching the surgical craft: from selection to certification. *Curr Probl Surg* 39: 573–659

Paper II

Larsen C R, Grantcharov T, Schouenborg L, Ottosen C, Sorensen J L and Ottesen B.
Objective assessment of surgical competence in gynaecological laparoscopy:
Development and validation of a procedure-specific rating scale.
BJOG. 2008 Jun;115(7):908-16.



Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale

CR Larsen,^a T Grantcharov,^b L Schouenborg,^a C Ottosen,^a JL Soerensen,^a B Ottesen^a

^aDepartment of Gynaecology and Obstetrics, The Juliane Marie Centre (for Children, Women and Reproduction), Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark ^bDepartment of Surgery, University of Toronto, St Michael's Hospital, Toronto, Ontario, Canada
Correspondence: Dr CR Larsen, Department of Gynaecology, The Juliane Marie Centre (for Children, Women and Reproduction), Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 3, DK-2100 Ø, Copenhagen, Denmark. Email crl@rh.regionh.dk

Accepted 27 February 2008.

Objective The purpose of this study was to develop a global- and a procedure-specific rating scale based on a well-validated generic model (objective structured assessment of technical skills) for assessment of technical skills in laparoscopic gynaecology. Furthermore, we aimed to investigate the construct validity and the interrater agreement (IRA) of the rating scale. We investigated both the gamma coefficient (Kendall's rank correlation), which is a measure of the strength of dependence between observations, and the kappa value for each of the ten individual items included in the rating scale.

Design Prospective cohort, observer-blinded study.

Setting Departments of Obstetrics and Gynaecology in Zealand, Denmark.

Population Twenty one gynaecologists or gynaecological trainees.

Material and methods Twenty-one video recordings of right side laparoscopic salpingectomies were collected prospectively, eight from novices (defined as <10 procedures), seven from intermediate experienced (20–50 procedures) and six from experts

(>200 procedures). All operations were performed by the same operative principles and using a standardised technique. The recordings were analysed by two independent, blinded observers.

Main outcome measures Construct validity of the rating scale based on operative performance (median of total score) and interrater reliability.

Results There were significant differences between the three groups: median score of novices 24.00 versus intermediate 29.50 versus expert 39.50, $P < 0.003$) The IRA was 0.83 overall. The gamma correlation coefficient was 0.91. The kappa values varied from 0.510–0.933 for each of the individual items of the rating scale.

Conclusions The procedure-specific rating scale for laparoscopic salpingectomy is a valid and reliable tool for assessment of technical skills in gynaecological laparoscopy.

Keywords Assessment, construct validity, gynaecology, interrater reliability, laparoscopy, salpingectomy.

Please cite this paper as: Larsen C, Grantcharov T, Schouenborg L, Ottosen C, Soerensen J, Ottesen B. Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale. BJOG 2008;115:908–916.

Introduction

Laparoscopy, training and assessment

An increasing number of gynaecological surgical procedures are presently performed by laparoscopic technique. This leads to an increasing demand for evidence- and proficiency-based education, training and assessment of laparoscopic skills. Traditionally, education has been based on the apprenticeship model leaving both training and assessment of knowledge as well as technical skills subjective and unstructured.¹ So far, little has been carried out to develop and integrate structured basic training in laparoscopy in the surgical curriculum, and

no validated structured system for objective assessment of technical surgical skills in gynaecological laparoscopy is available. Consequently, we need a feasible, structured and objective system for assessment of both technical and procedural skills.

In Denmark, 94% of all surgically treated ectopic pregnancies are managed by laparoscopy.² The laparoscopic salpingectomy was in the present study chosen as the operative procedure for developing a method for objective structured assessment of technical surgical skills based on human operations. Laparoscopic salpingectomy is a basic laparoscopic procedure possessing the necessary complexity for skills

assessment. Furthermore, salpingectomy is one of the first laparoscopic procedures a trainee is exposed to. Finally, all specialised gynaecologists and obstetricians working on call should be able to perform a laparoscopic salpingectomy.

Assessment systems

Several systems to assess technical skills in minimally invasive procedures have been developed; some based on error detection and some on rating the technical skills. The best described error detection systems is the Time–Error matrices by Seymour *et al.*^{3,4} and the Observational Clinical Human Reliability Assessment system^{5–7} (OCHRA) by Tang *et al.* The Time–Error matrix has later successfully been modified to a more detailed version,⁸ closer to the very complex OCHRA. Characteristic for both systems is that they are highly detailed, thereby very time-consuming and difficult to implement and use for the observers, resulting in low feasibility.

Objective structured assessment of technical skills

In the late 1990s, Martin *et al.*⁹ and Reznick *et al.*¹⁰ developed the approach for assessing technical skills called Objective Structured Assessment of Technical Skills (OSATS). The OSATS was developed on the basis of the Objective Structured Clinical Examination. OSATS was originally developed for bench station test and consists of a task-specific checklist and global rating scale (GRS). The GRS has seven items, each evaluated on a global 5-points Likert-like scale where the lowest, middle and highest scores are defined by explicit descriptions of performances.¹⁰

It is well documented that GRSs are reliable, have high interrater reliability (IRR) and construct validity.⁹ Since first presented, the OSATS has been modified and tested in many different surgical areas such as open surgery,¹¹ laparoscopic surgery,¹² vascular surgery,¹³ and microsurgery,¹⁴ urology,¹⁵ ophthalmology (Global Rating Assessment of Skills in Intraocular Surgery),¹⁶ gynaecology^{17,18} and obstetrics.^{19,20} Most tests were conducted using OSATS as a bench test, fewer in clinical set-up, and none of them for laparoscopic gynaecology. In the different studies referred to above, the Cronbach's coefficient alpha (expressing the internal consistency reliability) varies from 0.71⁹ to 0.97.²¹ Unfortunately, IRR (interrater agreement [IRA]), which expresses the proportion of times to which two or more independent observers agree absolutely on their rating of a subjects performance,²² is not always described in these studies. An overall agreement among two observers ≥ 0.8 is, based on expert opinion, considered acceptable for a test system.²² In those studies where the IRR is stated, it varies from 0.70⁹ to 0.97.²³ The solid evidence of the reliability, feasibility and construct validity of OSATS, and the modified versions for different specialties, have almost completed the OSATS as the gold standard in assessment of technical skills. The Toronto group creating

the OSATS also made a modified assessment system for evaluation video recordings of laparoscopic (gastrointestinal) surgery. The modified system was developed for operations on anaesthetised pigs, not human operations.²⁴ The system consists of a reduced GRS suitable for laparoscopic surgery and a task-specific rating scale called operative component rating scale (OCRS).

Objective

The purpose of this study was to develop, and investigate the construct validity, IRA and gamma correlation on a procedure-specific rating scale for objective structured assessment of technical surgical skills in laparoscopic salpingectomy.

Methods

After a hierarchical task analysis²⁵ of the laparoscopic salpingectomy, we constructed a modified rating scale for laparoscopic salpingectomy called Objective Structured Assessment of Laparoscopic Salpingectomy (OSA-LS; Table 1). The OSA-LS was based on both the original OSATS⁹ and the modified rating scale for laparoscopic cholecystectomy, developed by Grantcharov *et al.*¹² It consists of five general items and five task-specific items equivalent to the OCRS by Dath *et al.*²⁴ First, in a pilot study, the observers together assessed ten video recorded operations to standardise their assessments and to adjust the scale (data not shown). In the main study, 21 video recordings of right side laparoscopic salpingectomies were collected prospectively over a 6-month period, 8 performed by novices (defined by less than 10 procedures), 7 by intermediate experienced (20–50 procedures) and 6 by experts (>200 procedures). All the recorded operations were performed by different surgeons, but using a standardised operative technique (based on expert consensus) as described by Nezhat *et al.*²⁶ (Table 2). The two independent observers used the OSA-LS chart for assessment of the 21 unedited video recordings. The observers (L.S. and C.O.) were blinded for surgeon and proficiency group status. Both observers are experts in laparoscopic gynaecological surgery, having performed more than 2000 advanced laparoscopic procedures each. The introduction of Veress needle and placement of the trocars were not evaluated in this study.

Ethics

No additional ethical approval was needed according to the Danish National Committee on Biomedical Research Ethics.

Statistics

Data on the performance on each of the individual items included in the rating scale are ordinal. Ordinal data are categorical data where there is a logical ordering of the categories. The Likert-like scales that are also used in many surveys

Table 1. OSA-LS: assessment chart

	1	2	3	4	5
OSA-LS general skills					
Economy of movements	Many unnecessary movements		Efficient motion but some unnecessary movements		Maximum economy of movements
Confidence of movements: instrument handling	Repeatedly makes tentative or awkward moves with instruments		Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Economy of time	Too long time used to perform sufficiently		Intermediate time used to perform sufficiently		Minimal time used to perform sufficiently
Errors: respect for tissue	Frequently used unnecessary force on tissue or risk of damage by inappropriate use of instruments or instruments often out of sight		Careful handling of tissue but occasionally risk of (minimal) damage, or instruments out of sight		Consistently, handled tissues appropriately with no risk of damage, instruments always in sight
Flow of operation/operative technique	Imprecise, wrong technique in approaching the operative interventions, or constant supervisor corrections		Careful technique with occasional errors or little supervisor correction		Fluent, secure and correct technique in all stages of the operative procedure, no supervisor corrections
OSA-LS specific skills					
Presentation of anatomic structures	Poor retraction and exposure of fallopian tube and round ligament		Satisfactory retraction and exposure of fallopian tube and round ligament		Expert retraction and exposure of fallopian tube and round ligament
Use of diathermy	Using diathermy too close to healthy ovarian or other tissue, risk of damage		Mostly safe use minimal risk of damage		Perfectly safe use of diathermy, no risk of damage
Dissection of fallopian tube	Inadequate dissection of fallopian tube. Additional damage or bleeding or part of fallopian tube left <i>in situ</i>		Identified fallopian tube, adequate dissection little damage of other structures, little bleeding		Clearly identified fallopian tube, perfectly dissected no additional damage, no bleeding. Fallopian tube completely removed
Care for ovary, ovarian artery and pelvic wall	Using diathermy or cutting too close to ovarian artery high risk of bleeding or occlusion of vessel or cauterising ovary		Mostly safe use of instruments, low risk of arterial damage, little cauterising on ovary		Perfectly safe use instruments, no risk of cauterising or cutting the ovary, ovarian artery or other non-target tissue
Extraction of fallopian tube	Clumsily performed with major difficulty to catch the tissue, retract or get the tissue in the bag		Minor difficulty retracting or getting the tissue in the bag		Perfect retraction grasps end of structure or, easy placement of tube in bag
Σ General skills: _____ Σ Specific skills: _____ Σ All skills: _____ Total time in minutes: _____ Non-assessed items: _____					

is a typical example: 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree. Cumulated scores for subjects or groups of subjects are continuous data. Due to the sample size and the nature of the results, a Gaussian distribution could not be expected. Cumulated scores are presented as median and interquartile range (IQR) and compared using the Kruskal–Wallis nonparametrical comparison of mean. For *post hoc* analysis, Bonferroni corrected Mann–Whitney *U* tests were used. A *P* value of (two-tailed) <0.05 is

considered to be statistically significant. IRA was calculated as observation events agreements divided by the total number of observations for the single proficiency groups as well as for the entire sample. The gamma coefficient, a nonparametric rank correlation investigating agreement in ordinal categorical data, was used to investigate strength of correlations among the observers at single subject level. Systematic as well as nonsystematic disagreements were also analysed. Finally, to reveal items less agreeable, kappa values on single items level

Table 2. Operating instruction: laparoscopic salpingectomy, modified after Nezhat *et al.* (2000)²⁶

Salpingectomy: operation instruction (expert consensus)	
Instruments	Graspers, bipolar diathermy, scissors, rinse/suction, bag
Procedure	After introducing the trocars and pneumoperitoneum
1	Start the video recording
2	Insert your instruments, grasper in lateral trocar other instrument in medial trocar
3	Identify the anatomy
4	Operate from centre towards lateral
5	Use grasper in right hand and grasp the fallopian tube
6	Use bipolar grasper in left hand and use diathermy on salpinx and mesosalpinx
7	Start close to tubal corner of the uterus
8	Shift bipolar grasper to scissors in left trocar and cut the coagulated tissue close to fallopian tube
9	Continue alternated use of bipolar grasper and scissors to remove the fallopian tube. Use instruments in the trocars providing the most appropriate access to the tissue
10	Take care not to use diathermy on the ovary and the supplying artery and other non-target tissue
11	Use bag or gasper to remove the dissected tissue
12	Use rinse/suction device to clean up blood, use bipolar grasper to coagulate any remaining bleeding vessels/tissue
13	Pull out both instruments and tell the supervisor if you consider the operation performed. Stop video recording

are calculated. Analysis was performed using the Statistical Package for Social Sciences (SPSS®) 13.0 for Windows (SPSS Inc., Chicago, IL, USA). Graphics was made on Graphpad Prism 4.0 for windows (Graphpad Software Inc., San Diego, CA, USA).

Results

The independent and blinded evaluation of the operations demonstrated that the median score in the novice group was 24.00 (IQR 23.75–25.25), in the intermediate experienced group was 29.50 (IQR 28.00–31.00) and in the expert group was 39.50 (IQR 33.50–42.50). This revealed that the OSA-LS was construct valid and able to discriminate between all groups ($P < 0.03$) (Figure 1 and Table 3). The difference in overall score between novices and intermediate experienced gynaecological surgeons was 6 points and between intermediate experienced and experts 8 points. The overall IRA was 0.831, varying from 0.759 in the experienced group to 0.905 among the intermediate experienced gynaecological surgeons (Table 4).

The gamma correlation coefficient was 0.91 (95% CI 0.785–1.000) for all observations (Figure 2). The lowest correlation was found in the novice group, the highest correlation in the expert group. Even in the novice group, where the lowest gamma correlation coefficient was found, the discrepancy of the observers' ratings was randomly distributed. This emphasises that none of the observers systematically rated the performances differently from the other observer, i.e. neither more negative nor more positive.

The kappa value on items level (all 21 subjects) varied from 0.510 to 0.933, indicating that item 2, 4 and 6 were main sources of disagreement; in the other items, observers reached a higher degree of agreement (Table 5). The median time used for evaluating the unedited video recordings, including filling out the score table, was 16 minutes (range 7–35). Not all participating gynaecologists used an Endobag® (LiNA Medical UK Ltd, Dulford, UK) or other bag systems to remove the dissected tissue from the body. Some of the women underwent further surgery, e.g. hysterectomy, consequently the fallopian tube was removed en bloc at a later stage, together with additional dissected tissue. Item 10 has therefore been excluded from this validation study.

Discussion

Construct and discriminative validity

The OSA-LS for video evaluation of surgical skills in laparoscopic gynaecology was demonstrated to be feasible, had good construct validity and high IRA. Construct validity, a core property of a test, is the extent to which the test measures the trait that it purports to measure. The OSA-LS for video

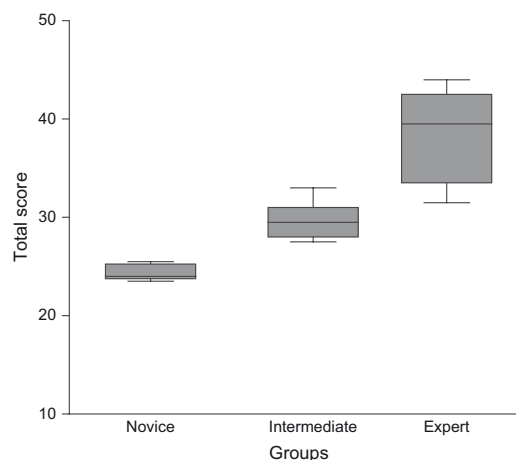


Figure 1. Box plot median score all groups, band represents median; boxes represents IQR; whiskers represents range.

Table 3. Construct (discriminative) validity of the OSA-LS: Comparison of median score of the three proficiency groups. Statistics: Kruskal–Wallis nonparametric, *post hoc* analysis: Mann–Whitney (Bonferroni corrected)

	<i>n</i>	Median	IQR		Range		<i>P</i> value	
			25% percentile	75% percentile	Maximum	Minimum	Kruskal–Wallis	Mann–Whitney (<i>post hoc</i>)
Novice	8	24.00	23.75	25.25	25.50	23.50	<0.001	Novice vs intermediate <0.03
Intermediate	7	29.50	28.00	31.00	33.00	28.00		Intermediate vs experts <0.03
Expert	6	39.50	33.50	42.50	44.00	31.50		
All	21	30.50	24.50	34.25	44.00	23.50		—

assessment in the operating room demonstrates construct validity, like the bench station OSATS did. Our results are consistent with the results found in other clinical specialties.^{9,10,12–14,16–18} The groups with different levels of experience were clearly discriminated by both observers using the OSA-LS. Based on these findings, it can be concluded that the test can be applied to test the laparoscopic abilities of trainees in obstetrics and gynaecology.

A major advantage of this OSA-LS system is that the assessment is based on a real human operation rather than a bench-simulator- or animal model, thereby testing how the surgeon is actually performing in operating room on humans rather than how they intend to perform the surgery by demonstrating the procedure in a model, simulator or animal.

According to the classic Miller's pyramid of competence development,²⁷ the 'Level 3: shows how' can be represented by evaluation of competence in the bench station test or simulator, whether the 'Level 4: Does' only can be represented by evaluation of competence in a real (human) operation. The OSA-LS provides us ability to test the competence at the 'Level 4'.

Another advantage of the OSA-LS is that it can serve as an excellent basis for structured feedback. Going through the operation video together with the OSA-LS evaluation could provide the trainee with valuable knowledge of his or her strengths and weaknesses and can potentially shorten the learning curve.²⁸

Table 4. Inter Rater Reliability at proficiency group level and at the overall level (bold)

Group	<i>n</i>	Items	Numbers of evaluated observations	Number of disagreement	IRR
Novice	8	9	72 (9 × 8)	14	0.806
Intermediate	7	9	63 (9 × 7)	6	0.905
Expert	6	9	54 (9 × 6)	13	0.759
Overall	21	9	189 (9 × 21)	32	0.831

Item 10 excluded, see text for details.

Nevertheless, the sample size in the present study is small, as in most of this kind of studies,^{9,20,21} consequently the results have to be interpreted with caution.

Performance range within the groups

Ideally, the range of performance should be narrow within each group, indicating that the subjects fit the group definition. In this study, however, we found a quite wide performance range in the expert group and a narrow performance range in the novice group. There could be several explanations for this. First, the figures could be coincidental, due to small sample size. More likely, the wide performance range could indicate that some experts are more skilled than others. The definition of the proficiency groups is traditionally

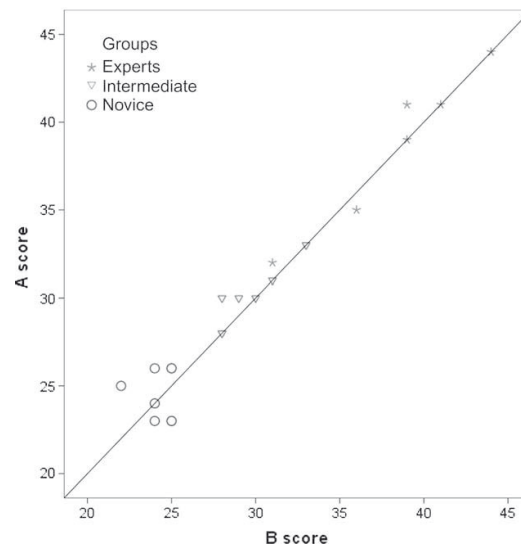
**Figure 2.** Scatter plot: observer B cumulative score for each individual (x axis) versus observer A cumulative score for same subjects (y axis). Identification line represents absolute agreement among observers. Distance from line represents magnitude of observer disagreement.

Table 5. Kappa values on single items level

Item	n	Kappa	SE	95% CI	
				Lower bound	Upper bound
1	21	0.851	0.100	0.655	1.000
2	21	0.653	0.132	0.394	0.912
3	21	0.933	0.065	0.806	1.000
4	21	0.625	0.129	0.372	0.878
5	21	0.789	0.112	0.569	1.000
6	21	0.510	0.149	0.218	0.802
7	21	0.731	0.147	0.443	1.000
8	21	0.695	0.137	0.426	0.964
9	21	0.695	0.131	0.438	0.952
10	0	—	—	—	—

carried out by number of surgical procedures performed or times spent in a given position. This is not the most appropriate way as the number of cases performed is not an objective measure of competency. Furthermore, individuals have different learning curves. Some individuals have innate abilities to perform laparoscopic surgery and a very steep learning curve, others may need a variable number of procedures to reach a plateau and some may never achieve proficiency due to poor neuropsychological abilities.²⁹ In fact, only the novices represent a truly defined proficiency group, quantitative as well as qualitative; having performed none or few procedures and being at the early stage of their learning curve. It is more difficult to define the intermediate and expert groups by a quantitative definition, such as number of procedures performed. They should, more accurately, until their true technical competence level is established objectively, be called 'quantitative experts' or 'experienced'. As a consequence of this, the educational system is currently replacing training and assessment systems that only counted the number of procedures performed by training and assessment systems based on competence levels.³⁰ Nevertheless, in most previous studies, experts have only been defined by number of performed procedures, thus leaving this problem as a challenge for future research.

The wider range in performance in the expert group can also be explained by a wider variety of cases operated by the expert group, a *case mix* situation. The novices will always get the easiest cases, and the experts the most complicated cases, such as cases with dense adhesions or significant pathological anatomy. This could influence the assessment. In the expert group, there was one low outlier, which was shown to be a more complicated case, due to a severe hydrosalpinx with dense adhesions to a cystic ovary. This is a weakness of all assessment systems developed from bench station tests, which were originally designed to standardised cases, because they

do not take into account that the same procedure in some cases is in reality more difficult in others. There are two ways to overcome this problem. Either the assessment system should only be used for cases of predefined complexity, for instance using only simple and uncomplicated cases for assessment purposes. Another way of dealing with the biological variation is by developing a graded system for case complexity. Multiplying the total score by a predefined factor according to case complexity might solve the problem. This, however, must be developed and validated in a separate prospective study.

Re-analysing the recordings used in this project, we believe that there are not only differences in level of difficulty of a given procedure, but there are also substantial differences in the performance levels among the experts. Additionally, compared with the novices, who all handled the tissue extremely gently, some of the experts seemed confident and fast, handling the tissue a bit roughly. This might not influence the surgical outcome but is detected in the OSA-LS system.

Talent selection

A continuing discussion among health authorities is whether assessment systems in surgery can be a method for recruitment and career guidance. Based on the figures from this study, and based on previous studies on simulator-based salpingectomy,^{31,32} this seems to be quite difficult. When the range is as narrow as seen in this investigation, it would be extremely difficult to distinguish talents from non-talents, unless more video recordings per individual were evaluated. It is, however, unknown how many video recordings would be needed to obtain valid information. This observation is consistent with findings in the original paper where it was stated that OSATS were not designed as a predictor of surgical skills for residents before entering specialty training.⁹

Interrater agreement

The level of agreement between two independent observers blinded for the test subjects training status is important in the evaluation of an assessment system. It reveals how unambiguous the test is, and thereby how valid it is if used by different independent raters. Several methods establishing the IRA are presented in the literature. Cohen's kappa value is often described as the best measurement for the degree of agreement among the observers. Fundamentally, kappa calculates the degree of agreement, but it also takes into account the degree of agreement that could be expected to occur by chance, hence argued to make this statistical test more robust (beyond-chance agreement). This is very important in a single item test, and when the outcome is binary, e.g. yes or no, where the agreement occurring by chance can be as high as 25%. In a multiple items test, like this modified ten items OSA-LS, evaluated on a five categories Likert-like scale, the

overall agreement occurring by chance is negligible. This makes the simple calculation of IRA (observation events agreements/total number of observation) a sufficient measure for general purpose. A disadvantage of using only kappa statistics (and IRA) is that kappa only gives a general value of the observer agreement. It does not make distinctions in-between various types and sources of disagreement. Besides, in a multi-item test like the OSA-LS presented, an IRA > 0.8 is perfectly acceptable but still leaves us with 20% disagreement and no information on where to find the disagreement. Consequently, we also investigated the correlations coefficient gamma and the kappa values at a single items level.

Gamma coefficient and kappa value

The gamma coefficient is a nonparametric rank correlation investigating agreement of ordinal categorical data. The gamma coefficient was used to investigate strength of correlations, or agreement, among the observers at single subject level. This test can reveal whether a possible disagreement is systematic or random among the observers, or if possible, disagreement is only found in certain individuals or groups of individuals. Values of the gamma coefficient range from -1, negative association to +1, perfect agreement; 0 indicates absence of association.³³ To explore that, we calculated the correlation of the observers to see whether a small discrepancy was systematic or nonsystematic and in which groups the correlation was highest. Figure 1 shows that the novice group subjects are very homogeneous in total score. However, this group also demonstrated the slightest correlation among the observers. In contrast, this disagreement is nonsystematic, leaving this group with the smallest dispersion. In this study, we found a very high overall correlation among the observers. As seen in the scatter plot, the correlation was higher for the intermediate and expert groups than for the novice group, but the discrepancy found in the latter was not systematic and thereby not influencing the total OSA-LS score for the individual subject. Based on these results, we conclude that the OSA-LS is suitable regarding correlation and systematic as well as random disagreement.

Looking at single items level, kappa values discovered some items more disagreeable than others. Items 2, 4 and 6 had the lowest kappa values, revealing quite some disagreement among the observers. This information could be used to discuss the interpretation of the item terms, and if necessary specify or rephrase the terms. If there is still a low degree of agreement, it should be considered to exclude the item from the list. However, in this study, the confidence intervals are large, most, except item 3, suggest that the level of agreement ranges from poor to excellent for each item making conclusions on kappa values difficult in a study of this size. Furthermore, it demonstrates that the simple IRA calculation probably is more suitable for this kind of observations, Likert-like scales with a small sample size.

Strengths and weaknesses using the rating scales

In contrary to the traditional method known from the apprenticeship educational model, the OSA-LS for laparoscopic gynaecology provides a valid, structured, objective and systematic method for assessment and feedback of technical skills. Besides valid and detailed assessment of the trainee, the global- and task-specific rating scales also provide clinical relevant information for constructive feedback. This is a great advantage compared with the use of laparoscopic simulators, box trainers and other metrics based systems for evaluation of skills. The simulator method although, is construct valid,^{31,32} but the metrics used in the assessment, like instrument angular path and instrument path length, have only relevance for improvement of dexterity, they do not have direct clinical implication. These parameters do therefore not provide clinically useful information for feedback on operative technique. Using the rating scales, the assessor can feedback the marks given for the different items to the trainee while going through the video recording, providing examples on where to improve. This could be carried out in a nonstressful setting and provide structured, objective and very detailed advice.

A disadvantage of the OSA-LS video evaluation is how time-consuming it is. However, it is an advantage that the evaluation is video based, giving the assessor the possibility to analyse the performance whenever it is convenient time wise. The fact that the evaluation can be blinded for surgeon is also a big advantage for objective assessment. The study has a possible bias by using the same experts for development of the OSA-LS scale and for the later validation of the rating scale. The internal consistency is high, based on the high IRR, while the external consistency has to be demonstrated by applying this rating scale in other departments and with other raters.

Certification

In certification matters, it is extremely important that a test is valid. Refusing a trainee with sufficient skills to operate will be inconvenient, to let a trainee surgeon with inadequate skills operate would be unacceptable. The high degree of construct validity and IRA makes it possible that this rating scale can serve as a basis in high stakes situations like certification and recertification. However, this study was not designed to define 'cut off level', this will still have to be based on expert consensus according to national surgical curricula.

Learning curves

Developing procedure-specific rating scales for different procedures also provide a good opportunity for assessment of learning curves for the different procedures. The learning curves combined with the advantages of feedback using the rating scales open the possibility to design high-quality training curricula in advanced laparoscopy.

Future studies

Using this validated OSA-LS scale for laparoscopic salpingectomy is of obvious interest to test the impact of simulator training of laparoscopic skills on real operations. Further investigations on the learning curve of individual trainees will also be of great interest. Finally, a second modification on the scale including a difficulty grading system to be applied on nonstandard cases should be developed.

Conclusion

The OSA-LS for assessment of technical surgical skills in laparoscopic gynaecology is construct and discriminative valid and has a sufficiently high degree of IRA and gamma correlation among independent observers blinded to surgeon. The kappa values at single items level revealed that seven of ten items represented a high IRA; only items 2, 4 and 6 needed better definitions or more training of the observers. The system can provide the trainee both objective assessment and detailed clinical relevant feedback.

Funding

Rigshospitalet, University Hospital of Copenhagen, funded the project.

Details of ethics approval

The Helsinki II declaration as well as local legalisations were respected, no additional ethical approval was needed according to the Danish National Committee on Biomedical Research Ethics; the video recordings were made anonymous, patients cannot be identified or tracked.

Contribution to authorship

C.R.L.: Principle investigator, design, acquisition of data, statistics, analysis and interpretation of data and drafting the paper. T.G.: Background research, design, results analysis and discussion and revising the paper. L.S.: Design, data collection and analysis and revising the paper. C.O.: Design, data collection and analysis and revising the paper. J.L.S.: Background research, design, statistics, discussion analysis and revising the paper. B.O.: Supervisor, fundraiser, design, results analysis and discussion and revising the paper.

Acknowledgements

Thanks to Svend Kreiner, Associate professor at Department of Biostatistics, Faculty of Health Sciences University of Copenhagen, for statistical support. Thanks to the departments of Gynaecology and Obstetrics at the Hospitals in

Gentofte, Herlev, Hillerød, Hvidovre, and Roskilde for collecting video recordings of operations. ■

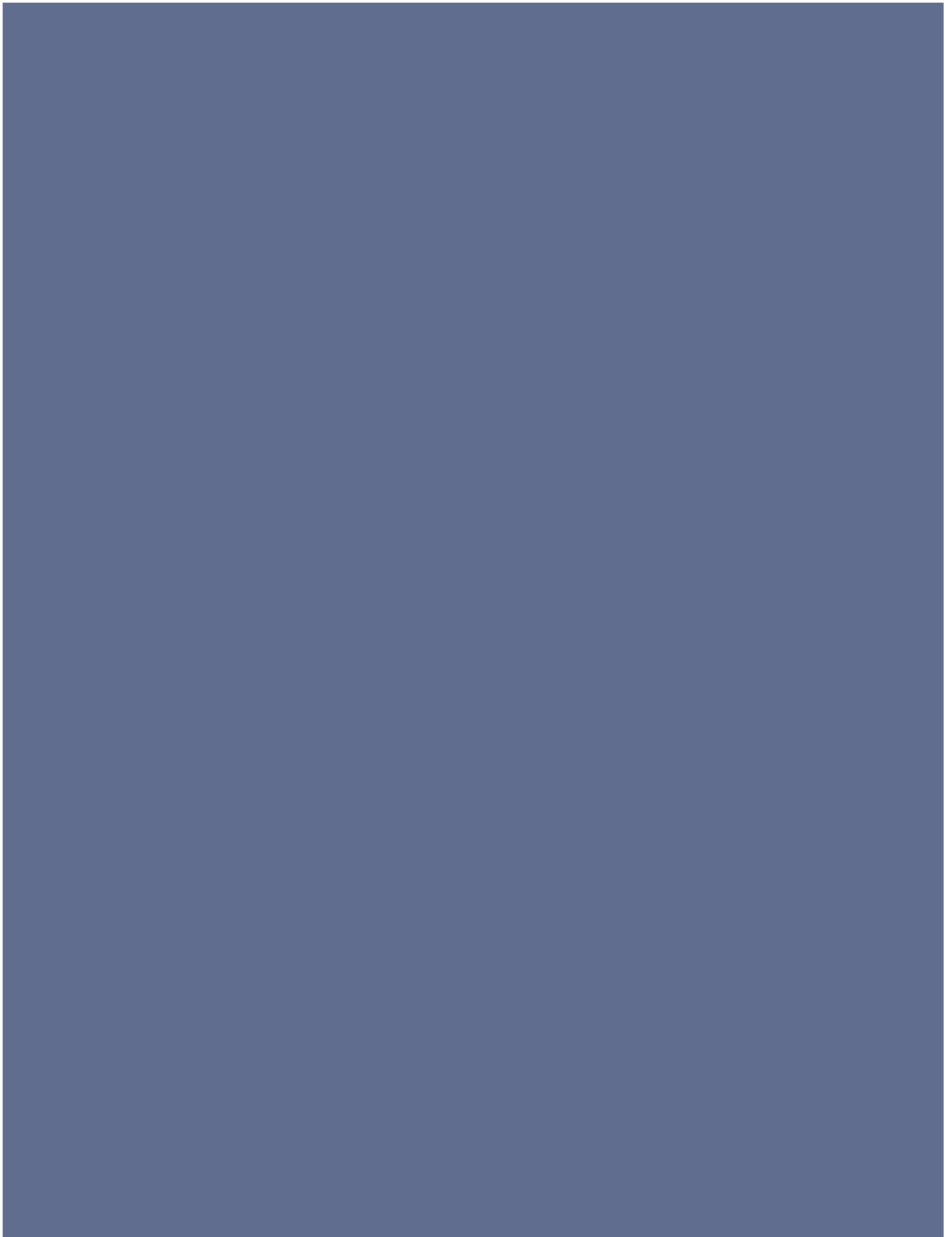
References

- 1 Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003;327:1032–7.
- 2 Lidegaard O, Hammerum MS. [The National Patient Registry as a tool for continuous production and quality control]. *Ugeskr Laeger* 2002; 164:4420–3.
- 3 Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Andersen DK, Satava RM. Analysis of errors in laparoscopic surgical procedures. *Surg Endosc* 2004;18:592–5.
- 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;236:458–63.
- 5 Tang B, Hanna GB, Cuschieri A. Analysis of errors enacted by surgical trainees during skills training courses. *Surgery* 2005;138:14–20.
- 6 Tang B, Hanna GB, Joice P, Cuschieri A. Identification and categorization of technical errors by Observational Clinical Human Reliability Assessment (OCHRA) during laparoscopic cholecystectomy. *Arch Surg* 2004;139:1215–20.
- 7 Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A. Competence assessment of laparoscopic operative and cognitive skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World J Surg* 2006; 30:527–34.
- 8 Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA III, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007;193:797–804.
- 9 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273–8.
- 10 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173:226–30.
- 11 Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *Am J Surg* 2006;192:372–8.
- 12 Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg* 2004;91:146–50.
- 13 Beard JD, Choksy S, Khan S; Vascular Society of Great Britain and Ireland. Assessment of operative competence during carotid endarterectomy. *Br J Surg* 2007;94:726–30.
- 14 Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS, et al. The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg* 2004;240:374–81.
- 15 Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol* 2002;167:1243–7.
- 16 Cremers SL, Lora AN, Ferrufino-Ponce ZK. Global Rating Assessment of Skills in Intraocular Surgery (GRASIS). *Ophthalmology* 2005;112: 1655–60.
- 17 Lentz GM, Mandel LS, Goff BA. A six-year study of surgical teaching and skills evaluation for obstetric/gynecologic residents in porcine and inanimate surgical models. *Am J Obstet Gynecol* 2005;193: 2056–61.

- 18 Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol* 2000;96:146–50.
- 19 Siddighi S, Kleeman SD, Baggish MS, Rooney CM, Pauls RN, Karram MM. Effects of an educational workshop on performance of fourth-degree perineal laceration repair. *Obstet Gynecol* 2007;109:289–94.
- 20 Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol* 2006;195:617–21.
- 21 VanBlaricom AL, Goff BA, Chinn M, Icasiano MM, Nielsen P, Mandel L. A new curriculum for hysteroscopy training as demonstrated by an objective structured assessment of technical skills (OSATS). *Am J Obstet Gynecol* 2005;193:1856–65.
- 22 Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003;17:1525–9.
- 23 Goff B, Mandel L, Lentz G, Vanblaricom A, Oelschlagel AM, Lee D, et al. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol* 2005;192:1331–8.
- 24 Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc* 2004;18:1800–4.
- 25 McLeod PJ, Steinert Y, Trudel J, Gottesman R. Seven principles for teaching procedural and technical skills. *Acad Med* 2001;76:1080.
- 26 Nezhat C, Siegler A, Nezhat F, Nezhat C, Seidman D, Luciano A. Operations on the Fallopian Tube. In: *Operative Gynecologic Laparoscopy: Principles and Techniques*. San Francisco, CA: McGraw-Hill; 2000. pp. 246–51.
- 27 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65 (9 Suppl):S63–7.
- 28 Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc* 2007;21:2240–3.
- 29 Grantcharov TP, Funch Jensen P. Learning curve patterns in technical skills acquisition in laparoscopic surgery. Can everybody learn it? Association of Surgeons of Great Britain and England Clinical Congress 3rd to 5th May 2006, Abstract no. 10745.
- 30 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;357:945–9.
- 31 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;20:1460–6.
- 32 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* 2006;113:1382–7.
- 33 Agresti A. *Analysis of Ordinal Categorical Data*. Hoboken, NJ, USA: Wiley, 1984.

Paper III

Larsen C R, Sorensen J L, Grantcharov T, Ottosen C, Schouenborg L, Dalsgaard T, Schroeder T V and Ottesen B.
Impact of Virtual Reality Training in Laparoscopic Surgery: A Randomised Controlled Trial.
In review; BMJ.



Effect of virtual reality training on laparoscopic surgery: randomised controlled trial

Christian R Larsen, clinical research fellow,¹ Jette L Soerensen, assistant professor and consultant,² Teodor P Grantcharov, assistant professor and consultant,³ Torur Dalsgaard, consultant,⁴ Lars Schouenborg, consultant,⁴ Christian Ottosen, consultant,⁴ Torben V Schroeder, professor and consultant,⁵ Bent S Ottesen, managing director and professor at the Juliane Marie Centre⁶

¹Department of Gynecology, Juliane Marie Centre for Children, Women and Reproduction, Copenhagen University Hospital Rigshospitalet, Blegdamsvej 9, DK-2100 OE, Copenhagen, Denmark

²Department of Obstetrics, Juliane Marie Centre, Copenhagen

³Division of General Surgery, St Michael's Hospital, Toronto, ON, Canada

⁴Department of Gynecology, Juliane Marie Centre, Copenhagen

⁵Department of Vascular Surgery, Abdominal Centre, Copenhagen University Hospital Rigshospitalet, Copenhagen

⁶Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark

Correspondence to: C R Larsen
crl@dadlnet.dk

Cite this as: *BMJ* 2009;338:b1802
doi:10.1136/bmj.b1802

ABSTRACT

Objective To assess the effect of virtual reality training on an actual laparoscopic operation.

Design Prospective randomised controlled and blinded trial.

Setting Seven gynaecological departments in the Zealand region of Denmark.

Participants 24 first and second year registrars specialising in gynaecology and obstetrics.

Interventions Proficiency based virtual reality simulator training in laparoscopic salpingectomy and standard clinical education (controls).

Main outcome measure The main outcome measure was technical performance assessed by two independent observers blinded to trainee and training status using a previously validated general and task specific rating scale. The secondary outcome measure was operation time in minutes.

Results The simulator trained group (n=11) reached a median total score of 33 points (interquartile range 32-36 points), equivalent to the experience gained after 20-50 laparoscopic procedures, whereas the control group (n=10) reached a median total score of 23 (22-27) points, equivalent to the experience gained from fewer than five procedures (P<0.001). The median total operation time in the simulator trained group was 12 minutes (interquartile range 10-14 minutes) and in the control group was 24 (20-29) minutes (P<0.001). The observers' inter-rater agreement was 0.79.

Conclusion Skills in laparoscopic surgery can be increased in a clinically relevant manner using proficiency based virtual reality simulator training. The performance level of novices was increased to that of intermediately experienced laparoscopists and operation time was halved. Simulator training should be considered before trainees carry out laparoscopic procedures.

Trial registration ClinicalTrials.gov NCT00311792.

INTRODUCTION

Laparoscopy has become the standard approach for many conditions in most surgical specialties.¹⁻³ This development has been driven by the desire for less surgical trauma, faster postoperative recovery, shorter

hospital stay, and better cosmetic results, and a sales drive by the medical industry.⁴ It is evident, however, that laparoscopy is associated with a longer operation time and a higher rate of surgical complications during the learning curve of the surgeons. This has been verified in many different specialties, including general^{5,6} urological,^{7,8} paediatric,⁹ and gynaecological surgery.¹⁰ The possibility of overcoming these problems during the learning curve by appropriate training and ensuring that surgeons perform a sufficient number of procedures has also been documented.¹¹ The technical skills needed for laparoscopic surgery are fundamentally different from those for traditional open surgery, leading to a prolonged learning curve. The primary obstacles in learning laparoscopy are psychomotor and perceptual. The unique nature of laparoscopic surgery combined with an increasing focus on patients' safety and rights, the present decrease in working hours, and concern over costs of operating theatre time are factors that challenge the traditional surgical approach and contribute to a growing need for novel methods in the training of laparoscopic surgeons.¹² Although virtual reality simulation has the potential to offer important advantages in the area of training for new skills and procedures, evidence on the transfer of skills from the simulated environment to the operating theatre is still limited.^{13,14}

We investigated the impact of training using a virtual reality simulator on the quality of skills acquired for a key gynaecological procedure. The investigation was carried out as a prospective, randomised, observer blinded, controlled trial, according to the guidelines of the consolidated standards of reporting trials (www.consort-statement.org).

METHODS

From September 2006 to June 2007, 32 trainees in gynaecological specialty training years 1 and 2 (post-graduate years 3-8; see box), with no experience of advanced laparoscopy (defined as all laparoscopic procedures involving coordination of more than one instrument), were included in the study. Of a total cohort of 42 (38 women and four men) trainees in the

region, eight were not eligible, as they were too experienced and four came from the two departments that did not participate in the trial. Of the remaining 30 eligible trainees, the first 24 who volunteered were enrolled. They came from seven of nine gynaecology departments in the Zeeland region of Denmark (population 2.3 million): Gentofte hospital (five trainees), Herlev hospital (n=4), Roskilde hospital (n=4), Hillerød hospital (n=1), Holbaek hospital (n=1), Hvidovre hospital (n=2), and Rigshospitalet hospital (n=7).

Randomisation and blinding

To ensure that the trainees' baseline characteristics were similar within and between each group, we chose a stratified randomisation based on previous experience of simple laparoscopy (defined as laparoscopic procedures performed using a single instrument, such as diagnostic laparoscopic sterilisation (clips) or diagnostic laparoscopy). The Clinical Trial Unit at Copenhagen University independently randomised the trainees by computer to intervention or control groups. The randomisation procedure was concealed and achieved by using the trainees' unique personal identification number (central personal register number).

Trainees in the intervention group were given an oral introduction to the simulator and the rating scale used for outcome measure. Any operations done during the study were recorded. Owing to the nature of the trial it was not possible to blind the trainees to their allocated group, but all involved departments, supervisors, and staff in the operating theatres were blinded to the trainee's group, and the assessors of outcome were blinded to both the trainee and their allocated group. The primary investigator saw the data only after completion of all assessments and once data had been loaded in a database.

The control group was to continue standard clinical education. During the study no trainee in either group was allowed to perform advanced laparoscopic surgery, only simple laparoscopy or to assist senior colleagues. To check that randomisation had been successful, the control group were trained in the simulator after the trial. Their baseline performances were indistinguishable from those of the intervention group.

Equipment

The virtual reality laparoscopy simulator program (LapSim Gyn v 3.0.1; Surgical Science, Gothenburg, Sweden) was run on an IBM T42 computer in a docking station (Pentium M 1.8 GHz/512 MB RAM; IBM,

Armonk, NY, USA) using an interface with a diathermy pedal (Virtual Laparoscopic Interface; Immersion, San Jose, CA, USA). The operations took place in the operating theatres of the participating departments and were recorded on DVD using a camera attached to the laparoscope for later blinded evaluation. During the operation one of the authors (CRL or designated TD) observed the procedure to record the handling of the surgical instruments, any involvement of the supervisor, whether the standard procedure for the operation was followed, and whether the recording was done correctly, finalised, and assessed.

Simulator training

The intervention group undertook a specific training programme in the simulator. The programme comprised two parts: firstly, training in the two basic skills of "lifting and grasping" and "cutting" during which the trainees were introduced to the simulator environment and the different instruments; secondly, one procedure specific task in which the trainee had to carry out a complete right sided salpingectomy while preserving the ovary. The training in basic skills was done once in each training cycle of 45-60 minutes and the salpingectomy repeated continually during the remainder of the cycle. The simulator provided the trainees with instant feedback on time, path length and angular path of the instruments' movements, bleeding, cutting of uncoagulated arteries, and use of diathermy on non-target tissue. The training sessions were repeated until the expert criterion level was reached in two consecutive and independent simulations. The proficiency criteria were established by experts in previous studies of construct validity and learning curves.^{15 16} The requirements and settings of the simulator are available at www.skopisimulator.rh.dk.

Surgical procedure

The trainees performed their first salpingectomy at their local gynaecological department and were supervised by a senior colleague who was told about the purpose of the trial. To make comparison of performance easier, the trainees all carried out procedures on the right side. The patients were admitted for elective salpingectomy before treatment for infertility or for prophylactic removal of fallopian tubes and ovaries owing to a positive test result for breast cancer gene 1 (BRCA1). The trainees were not allowed to operate on patients who had undergone previous open or laparoscopic surgery below umbilicus, had possible abdominal malignant disease, had an American Surgical Association score ≥ 3 (patients with severe systemic disease), had a body mass index less than 18 or more than 27, had haemophilia, or had other factors of potential influence on the surgical procedure. The operations followed a modified standard procedure on the basis of expert consensus.^{17 18} The supervisors were allowed to give oral instructions only, and one researcher was present to observe the procedure and to record who handled the instruments.

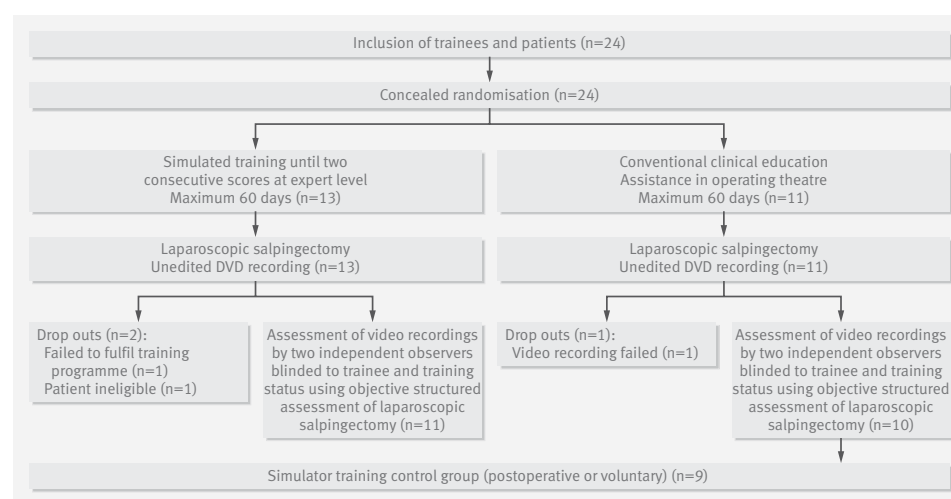
Duration of specialist training in Denmark

Preregistration house year—one year

Introduction before specialist training—one year

Optional additional training to qualify for further specialisation (possibility for writing PhD thesis)—1-3 years

Specialist training—four years



Flow of trainees through trial

Outcome measures

The primary outcome measure was technical performance, measured as total score (10-50 points) using the objective structured assessment of laparoscopic salpingectomy, which comprises a five item general rating scale and five item task specific rating scale.¹⁹ Two independent observers blinded to trainee and allocated group assessed the recorded operations. The secondary outcome measure was operating time in minutes. The reliability of the assessment was determined by calculating the inter-rater agreement (number of agreements for each of the assessed items divided by total number of assessed items) and the γ coefficient. We present outcomes as medians and inter-quartile ranges.

Power calculation

The power calculation was based on a previous validation study on the procedure specific scale of the objective structured assessment of laparoscopic salpingectomy.¹⁹ This study showed a difference of six points between novice laparoscopists (0-5 procedures) and intermediately experienced laparoscopists (30-50 procedures). An improvement of skills to the level of 30 or more points was considered acceptable. On the basis of these findings we chose the minimal relevant difference to be six points. We determined that with an α of 0.05

(two sided) and a power of 80% ($\beta=0.2$ giving $Z_{\alpha}=1.96$ and $Z_{\beta}=0.84$, largest $SD=4.40$) we required 18 or more trainees. To compensate for possible drop outs, we added an additional third to the 18, totalling 24 trainees.

Statistical analysis

We present cumulated scores as medians (average score of two observers), compared using non-parametrical analysis (Mann-Whitney U test). We considered a two tailed P value of 0.05 or less to be statistically significant and an inter-rater agreement and γ coefficient of 0.8 or more for each to be acceptable. Analysis was done using SPSS 13.0 for Windows. Graphics were made using Graph pad Prism (Graph Pad, San Diego, CA, USA).

RESULTS

Eight of the total cohort of 42 trainees (38 women, four men) were ineligible for the study as they were too experienced and four came from the two departments not participating in the trial. The remaining 30 trainees agreed to participate. The first 24 were enrolled; 22 (90%) were women, representing the current sex distribution among trainees in obstetrics and gynaecology in Denmark (figure). The average age of the trainees was 32.8 years (range 26-42 years), and 23 were right handed. Eleven trainees were randomised to virtual reality training in laparoscopic salpingectomy and 10 were randomised to traditional clinical education. Table 1 shows the baseline characteristics of the trainees. Two trainees were subsequently excluded from the simulator trained group because one failed to complete the training programme and the other was involved in an operation that was cancelled because of anatomical abnormality and suspected malignant disease in the patient. One trainee was excluded from the control group because of a technical fault in the DVD recorder used to record the operation.

Table 1 | Baseline characteristics of gynaecology trainees randomised to virtual reality simulator training in laparoscopic salpingectomy or to standard clinical education (controls). Values are numbers of trainees unless stated otherwise

Characteristics	Simulator trained group (n=13)	Control group (n=11)
Men	1	1
Women	12	10
Mean (range) age (years)	33.3 (30-42)	32.4 (26-38)
Experience of simple laparoscopy	6	5
No experience of simple laparoscopy	7	6

RESEARCH

Table 2 | Impact of virtual reality simulator training on surgical performance and operation time. Values are medians (ranges; interquartile ranges) unless stated otherwise

Outcome measure	Simulator trained group (n=11)	Control group (n=10)	P value*
Surgical performance:			
Total score (points)	33 (25-39; 32-36)	23 (21-28; 22-27)	<0.001
% reaching ≥30 points	82	0	
Operation time:			
Total time (minutes)	12 (6-24; 10-14)	24 (14-38; 20-29)	<0.001

Inter-rater agreement 0.79. γ -coefficient 0.83 (95% confidence interval 0.68 to 0.98).

*Mann-Whitney U test.

The median total score on the general and task specific rating scale reached 33 points (interquartile range 32-36 points) in the simulator trained group and 23 (22-27 points) in the control group ($P<0.001$, table 2).

The median total time to complete the procedure was 12 minutes (interquartile range 10-14 minutes) in the simulator trained group compared with 24 (20-29 minutes) in the control group ($P<0.001$, table 2). Twenty one operations were assessed.

The median number of simulated salpingectomies needed to reach the proficiency level in the simulator trained group was 28 (24-32 salpingectomies). The control group was offered simulator training after the study operation; nine of the 11 trainees in this group volunteered and a median of 26 (23-32) simulated operations were needed to reach the proficiency level ($P=0.70$). The mean time spent on training using the simulator was 7 hours and 15 minutes (5h 30 min-8h 0 min) in the intervention group and 7 hours and 0 min (5h 15 min-7h 45 min) in the control group ($P=0.65$; table 3). The baseline score (first attempt) was 8 (5-15) in the simulator trained group and 9 (7-19) in the control group after training ($P=0.70$; table 3). All trends of differences in baseline characteristics were not statistically significant.

The time used by the assessors to fill in the rating chart was the mean total operation time plus five minutes for each DVD recording. The inter-rater agreement was 0.79 (166/210). The γ coefficient used to investigate strength of correlations among the observers at single subject level reached 0.83 (95% confidence interval 0.69 to 0.99).

DISCUSSION

Proficiency based virtual reality training in laparoscopic salpingectomy compared with standard clinical education was associated with a clinically important

improvement of operative skills during the actual procedure. The learning curve in the operating theatre was also shorter. On the rating scale used in this study, which had previously been validated in a separate investigation, novices (fewer than five procedures) scored a median 24 points, and intermediately experienced trainees (20-50 procedures) a median 33 points compared with a median 39 points for experts.¹⁹ The clinical implications of the present findings are thus extensive. After training in a specific procedure to a predefined (proficiency based) level inexperienced trainees progressed from the performance level of a novice to that of an intermediately experienced gynaecologist, assessed in their first complex laparoscopic procedure. By using simulator training it might be possible to bypass the early learning curve, which is known to be associated with an increased rate of complications.²⁰ This study was not designed to investigate complication rates, and conclusions in this area must be drawn cautiously. In general it is difficult to use patient outcomes to evaluate a medical training course. Firstly, in contrast with trials of a single intervention (for example, a new drug) medical education is a complex intervention involving many interconnecting parts and different layers.²¹ Secondly, assessment of surgical technical skills of individual trainees will need to be based on surrogate end points rather than outcomes such as morbidity or mortality because it is an ethical imperative that an operation performed by a supervised novice ought to have the same outcome as that of the supervisor. Training may cost time and some inconvenience for the patient but should never jeopardise safety or outcome. Thirdly, to show differences in outcome, based on a training course, the numbers of trainees should by far outnumber the total number of trainees available, thus making such a trial unfeasible.

Operating time

Although operating time might be greater with novice surgeons, the outcomes of a supervised operation ought to be the same. The time to complete the laparoscopic salpingectomy was reduced by half. As the operating theatre serves both productivity and educational purposes, shorter operation times are of benefit.

The present results emphasise that by using virtual reality simulator training the surgical community can meet the need for proficiency based basic training in laparoscopy. These results also show that criterion based procedural training using a virtual reality simulator can help compensate for reduced working hours by bringing trainees to a higher level of performance more quickly. Traditional training depends on the supply of suitable procedures for training purposes, whereas simulator training can be used according to demand. To achieve an average of 28 salpingectomies can take a year or more in clinical practice, compared with eight hours of intensive training using the simulator.

Finally, reducing the operating time by half, from 24 minutes in the control group to 12 minutes in the

Table 3 | Number of sessions and duration of training in virtual reality simulator training programme in intervention group before training and in control group after surgery

Variable	Simulator trained group (n=11)	Control group (n=9)*	P value†
No (range) of training sessions	28 (16-39)	6 (19-43)	0.76
Duration (range) of training	7h 15m (4h 15m-9h 30m)	7h 0m (4h 0m-9h 15m)	0.70
Median (range) score on first attempt (%)	8 (5-15)	9 (7-19)	—

*Voluntary simulator training after surgery.

†Mann-Whitney U test.

simulator trained group, might not seem important. Nevertheless it has been estimated that during residency the additional costs of longer operating time is about (1997) \$48 000 (£31 841.00; €35 907.00) per graduate.²² For more extensive surgery or just a bilateral procedure, increased use of simulator training could reduce novice operating time substantially, easing the pressure on the often limited capacity for training in the operating theatre.

Transfer of skills

Despite virtual reality simulators being introduced more than a decade ago, evidence on the effect of simulators on performance in the operating theatre is still sparse.¹³ To date no published studies on the transfer of technical surgical skills from simulator to real operations had exceeded grade 2a evidence; in our study the level of evidence is 1b. These findings are supported in a systematic review,¹⁴ which also showed that the randomised controlled trials on effect of simulator training were generally of poor scientific quality. The major problems were randomisation with inadequate concealment of allocation, lack of power calculations or insufficient sample sizes, lack of observer blinding, insufficient training time or goals, non-standardised comparators in control groups, and disparate interventions in intervention groups. Most studies also used surrogate end points such as bench station or simulator tests to establish the transferability of skills, instead of assessing skills in real operations. The conclusion of the meta-analysis was that only few studies possess the necessary quality, that two studies showed a positive effect^{23,24} (real operation) and one study no effect²⁵ (simulated operation) of simulator based training. Another common feature of the previous studies is that they were all carried out using basic skills rather than procedure specific simulation. We used a procedural simulator, which provides training in both psychomotor and cognitive skills, thereby also improving knowledge of the surgical procedure and its potential pitfalls. There are probably several reasons for the significant impact on both performance and time in our study compared with previously published studies. Firstly, the simulator provides a realistic graphic presentation of anatomy in the surgical field and a good feedback system enabling the trainees to adjust their performance. Secondly, using predefined training goals (the expert proficiency level) instead of fixed time or numbers, encouraged the trainees to rehearse until they reached the maximal effect of the simulator training. Thirdly, instead of selecting medical students who might not choose a surgical career, we studied highly motivated trainees who needed to learn laparoscopic skills for the future.

We measured the impact of simulator training on a salpingectomy, which is an important gynaecological laparoscopic procedure and can be viewed as a key operation possessing all the core skills needed for most laparoscopic procedures. We did not test external validity and reproducibility beyond the specialty of

gynaecology. However the skills that were improved can probably be considered as genuine core skills, which would be beneficial across the surgical specialties.

The flow of an operation is based on both technical skills and knowledge of the procedure, and therefore we expect the impact of simulator training would apply to other laparoscopic procedures, although the effect would be more visible on the technical side than in the knowledge of procedural skills. These suppositions are supported by a contemporary Swedish study on procedural virtual reality simulator training of cholecystectomy, which reached conclusions similar to those of the present study.²⁶

Observer reliability

In the present study the γ coefficient showed that there was no systematic discrepancy among the raters. The inter-rater agreement also reached a sufficient level, which is evidence of a valid and reliable assessment.

The investigation was carried out in the same way that a curriculum integrated training course most likely would be implemented. The internal consistency of the trial could have been higher if all the trainees had operated in the same theatre, using the same technical equipment, and with the same supervisor and staff. However, by showing the effects of simulator training in settings closely resembling a regional simulator training course the external validity was improved. The primary investigator helped the trainee to use the simulator and introduced the different training modules but did not teach laparoscopic techniques. The feedback on performance was based on assessment in the simulator. A designated supporter at the training session could, however, be a source of bias. A setting where trainees practise by themselves could eliminate this potential source of bias. Finally, performing laparoscopic surgery also consists of identifying diseased anatomy, communication, teamwork, decision making,²⁷ and leadership, alternative plans, and conversion to open surgery if needed.^{28,29} These non-technical skills are taught in the currently existing virtual reality systems to a limited degree only. In this study we did not provide training in these non-technical skills or assess them. Simulator training should probably be considered only as a supplement or preoperative training; further education and practice in the operating theatre as well as further development of more complex virtual surgical environments (hybrid simulation)³⁰ is still required.

Conclusion

It is possible to transfer skills acquired during proficiency based training using a virtual reality simulator to a real operation. Training in proficiency based skills should be incorporated in a comprehensive surgical training and assessment curriculum for all residents before they operate on real patients. This can potentially improve patients' safety and improve efficiency in the operating theatre.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The European Working Time Directive has put extra pressure on surgical training programmes

Virtual reality simulators could contribute to the training of core skills for laparoscopy

High grade evidence of the effect of virtual reality simulator training on real operations is sparse

WHAT THIS STUDY ADDS

Training using a virtual reality simulator improved performance in a laparoscopic procedure

We thank Jørn Wetterslev at the Copenhagen Trials Unit for critical revision of the research protocol, assistance with the power calculation, and the concealed computer based randomisation of the participants.

Contributors: CRL (principal investigator) acquired the data, drafted the paper, did the statistical analysis, and obtained funding. TPG and JLS provided administrative support and critically revised the manuscript. JLS had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. TD acquired the data, provided technical support, and critically revised the manuscript. LS and CO acquired the data, provided administrative support, and critically revised the manuscript. TVS and BSO provided administrative support and supervised the study, critically revised the manuscript, and obtained funding. All authors conceived and designed the study and analysed and interpreted the data.

Funding: This project was supported by Copenhagen University Rigshospitalet Hospital. Trygfondet supplied various materials including computer hardware. Det Calssenske Fidecommis' Jubilaumsfond provided travel expenses. Aase and Ejner Danielsens foundation provided software maintenance and updates, DVD recorders, and a TV monitor. The Danish Society for the Protection of Laboratory Animals provided computer hardware and software. All phases of the present work including design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, and approval of the final manuscript were done independent of the funders.

Competing interests: None declared.

Ethical approval: The investigation fully complied with the Helsinki II declaration on biomedical research. The study was approved by the Danish National Committee on Biomedical Research Ethics (approval code (KF) 01 28 37 56). All study participants and patients were provided with written study documentation and were included in the trial after informed consent. The Danish Data Protection Agency approved the collection, analysis, and storage of the DVD recordings (approval code: 2005-41-5817).

- 1 Bruhat MA, Pouly JL. Endoscopic treatment of ectopic pregnancies. *Curr Opin Obstet Gynecol* 1993;5:260-6.
- 2 Keus F, Broeders IA, van Laarhoven CJ. Gallstone disease: surgical aspects of symptomatic cholecystolithiasis and acute cholecystitis. *Best Pract Res Clin Gastroenterol* 2006;20:1031-51.
- 3 Rosenmuller M, Haapamaki MM, Nordin P, Stenlund H, Nilsson E. Cholecystectomy in Sweden 2000-2003: a nationwide study on procedures, patient characteristics, and mortality. *BMC Gastroenterol* 2007;7:35.
- 4 Peters JD. Cutting the legal risks of laparoscopy. *OBG Management* 2002;14(10):47-55.
- 5 Karvonen J, Gullichsen R, Laine S, Salminen P, Gronroos JM. Bile duct injuries during laparoscopic cholecystectomy: primary and long-term results from a single institution. *Surg Endosc* 2007;21:1069-73.
- 6 Avital S, Hermon H, Greenberg R, Karin E, Skomick Y. Learning curve in laparoscopic colorectal surgery: our first 100 patients. *Isr Med Assoc J* 2006;8:683-6.
- 7 Kumar U, Gill IS. Learning curve in human laparoscopic surgery. *Curr Urol Rep* 2006;7:120-4.

- 8 Eto M, Harano M, Koga H, Tanaka M, Naito S. Clinical outcomes and learning curve of a laparoscopic adrenalectomy in 103 consecutive cases at a single institute. *Int J Urol* 2006;13:671-6.
- 9 Adibe OO, Nichol PF, Flake AW, Mattel P. Comparison of outcomes after laparoscopic and open pyloromyotomy at a high-volume pediatric teaching hospital. *J Pediatr Surg* 2006;41:1676-8.
- 10 Fleisch MC, Newton J, Steinmetz I, Whitehair J, Hallum A, Hatch KD. Learning and teaching advanced laparoscopic procedures: do alternating trainees impair a laparoscopic surgeon's learning curve? *J Minim Invasive Gynecol* 2007;14:293-9.
- 11 Carlsson S, Nilsson A, Wiklund PN. Postoperative urinary continence after robot-assisted laparoscopic radical prostatectomy. *Scand J Urol Nephrol* 2006;40:103-7.
- 12 Grantcharov TP, Reznick RK. Teaching procedural skills. *BMJ* 2008;336:1129-31.
- 13 Issenberg SB, McGaghie WC, Petrusa ER, Lee GD, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27:10-28.
- 14 Sutherland LM, Middleton PF, Anthony A, Hamdorf J, Cregan P, Scott D, et al. Surgical simulation: a systematic review. *Ann Surg* 2006;243:291-300.
- 15 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;20:1460-6.
- 16 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. *Br J Obstet Gynaecol* 2006;113:1382-7.
- 17 Nezhad C, Siegler A, Nezhad N, Nezhad Ce, Seidman D, Luciano A. Operations on the fallopian tube. *Operative gynecologic laparoscopy: principles and techniques*. San Francisco: McGraw-Hill, 2000:246-51.
- 18 Garry R. Laparoscopic surgery. *Best Pract Res Clin Obstet Gynaecol* 2006;20:89-104.
- 19 Larsen CR, Grantcharov TP, Schouenborg L, Soerensen JL, Ottosen C, Ottesen BS. Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure specific rating scale. *Br J Obstet Gynaecol* 2008;115:908-16.
- 20 Southern Surgeons Club. A prospective analysis of 1518 laparoscopic cholecystectomies. The Southern Surgeons Club. *N Engl J Med* 1991;324:1073-8.
- 21 Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694-6.
- 22 Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. *Am J Surg* 1999;177:28-32.
- 23 Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg* 2004;91:146-50.
- 24 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;236:458-63.
- 25 Ahlberg G, Heikkinen T, Iselius L, Leijonmarck CE, Rutqvist J, Arvidsson D. Does training in a virtual reality simulator improve surgical performance? *Surg Endosc* 2002;16:126-9.
- 26 Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McCluskey DA III, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007;193:797-804.
- 27 Kneebone RL, Nestel D, Chrzanowska J, Bamet AE, Darzi A. Innovative training for new surgical roles—the place of evaluation. *Med Educ* 2006;40:987-94.
- 28 Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003;327:1032-7.
- 29 Lingard L, Reznick R, DeVito I, Espin S. Forming professional identities on the health care team: discursive constructions of the 'other' in the operating room. *Med Educ* 2002;36:728-34.
- 30 Kneebone R. Simulation in surgical training: educational issues and practical implications. *Med Educ* 2003;37:267-77.

Accepted: 21 January 2009