

University of Groningen

Education in laparoscopic surgery

Kramp, Kelvin Harvey

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Kramp, K. H. (2016). *Education in laparoscopic surgery: All eyes towards in vivo training*. [Thesis fully internal (DIV), University of Groningen]. Rijksuniversiteit Groningen.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.



rijksuniversiteit
groningen

Education in laparoscopic surgery

All eyes towards in vivo training

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

woensdag 7 december 2016 om 12:45 uur

door

Kelvin Harvey Kramp

geboren op 31-10-1984
te Rotterdam

Promotor(es)

Promotor: Prof. dr. J.P.E.N. Pierie

Copromotor(es)

Copromotor: Dr. M.J. van Det

Beoordelingscommissie

Prof. dr. mw. A.D.C. Jaarsma

Prof. dr. E. Heineman

Prof. dr. J.F. Lange

Paranimfen

J.P. Entingh

M. Jones

The printing of this thesis is sponsored by:

- **Chirurgencoöperatie Oost-Nederland Hengelo-Almelo-Hardenberg**
- **Adriaan Metius Stichting**
- **Medisch Centrum Leeuwarden Academie**

ISBN: 978-90-367-9322-3

Cover design and illustrations by Sina Kazemian

Content

Chapter 1	General introduction and outline of the thesis	6
-----------	--	---

Part I: Aptitude

Chapter 2	The predictive value of aptitude assessment in laparoscopic surgery: a meta-analysis <i>Based on Medical Education 2016;50(4):409-427</i>	17
-----------	--	----

Part II: Training

Chapter 3	The Pareto-analysis for establishing content criteria in surgical training <i>Journal of Surgical Education 2016;73(5):892-901</i>	45
Chapter 4	Ergonomic assessment of the French and American position for laparoscopic cholecystectomy in the MIS Suite <i>Surgical Endoscopy 2014;28(5):1571-1578</i>	62
Chapter 5	Development of a standardized training course for laparoscopic procedures using a Delphi methodology <i>Journal of Surgical Education 2014;71(6):810-816</i>	77

Part III: Assessment

Chapter 6	Estimating the inter-rater reliability of surgical skills assessment <i>Submitted</i>	93
Chapter 7	Validity and reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy <i>Journal of Surgical Education 2015;72(2):351-358</i>	108
Chapter 8	Validity, reliability and support for implementation of independence-scaled procedural assessment in laparoscopic surgery <i>Surgical Endoscopy 2016;30(6):2288-2300</i>	122
Chapter 9	General discussion and future perspectives	145
Chapter 10	Summary	152

Chapter 11	Samenvatting	156
	Appendix	160
	Dankwoord	175
	Curriculum Vitae	177
	List of publications	178
	List of congress presentations	180

Chapter 1

General introduction and outline of thesis

General introduction

“Learning and teaching are fundamental, implicitly or explicitly, to human adaptation, socialization, culture change, and, at the broadest level, the production and reproduction of culture and society.”

Catherine Pelissier¹

Learning laparoscopic surgery

Learning curve

Learning, the adaptation or change of a system in response to stimuli, is vital for the existence of humans and the social constructions in which they participate. The first scientific publication on learning and memory dates back more than a century ago. In 1885, the German psychologist Ebbinghaus practiced the memorization of nonsense syllables and was the first to identify a learning curve.² He observed that the number of nonsense syllables he could remember increased substantially in the first attempts and that the improvements decreased in size as the number of attempts increased (Figure 1). Essentially, practice leads to cognitive changes that make the performance less effortful, faster and lead to less error. The by Ebbinghaus observed cognitive and psychomotor changes that occur as a response to task repetition have by some authors been referred to as one of the “biggest regularities” in human behaviour.³

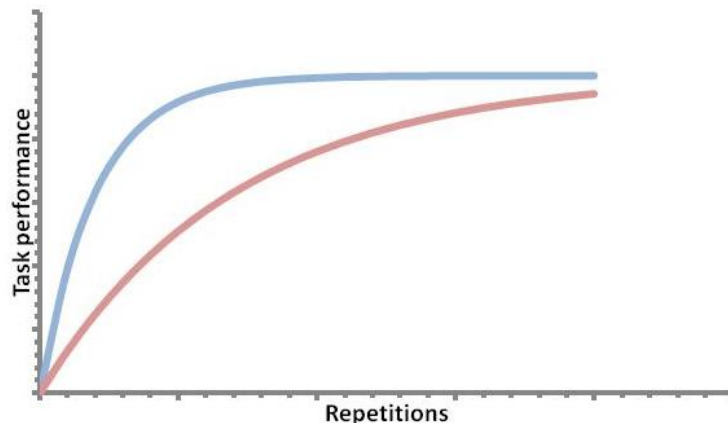


Figure 1: Learning curves. Blue: task with low difficulty, Red: task with high difficulty.

There are 3 different characteristics that can be distinguished in a learning curve: 1) the baseline performance; the initial performance level which can be influenced by previous experience in similar tasks, 2) the speed of skill acquisition; the shape or ‘steepness’ of the learning curve which is defined by the learning rate and 3) the learning plateau; the asymptotic part of the learning curve that is primarily defined by external factors. The characteristic phases of learning are the same in simple tasks (e.g. pushing a button) as in more complex tasks (e.g. surgery).^{3,4}

The learning curve can mathematically be expressed as performance level as a function of the number of repetitions of the task in a power law^{3,4}:

$$\text{Performance level} = P - [(P - B) \times N^{-\alpha}]$$

,wherein P = learning plateau, B = baseline performance, N = the number of repetitions, and α is the learning rate. Some state that the learning curve expressed with a power law is an artefact of

averaging the data of multiple people from multiple series and that the learning curve of an individual can better be described with an exponential law.³ Yet, the use of an exponential law instead of a power law does not change the above described basic characteristics of the learning curve.

Learning in surgery

Surgical trainees enter the operating room (OR) with the intention of learning as much as possible through an experience that has been highly organized to provide the best available patient care. It is through this experience that the trainee is expected to learn the skills to become a respected member of the OR team. These skills range from the technical skills needed to safely perform surgical procedures on the living human body⁵ to the understanding of the unique roles and responsibilities of the different members of the OR team.⁶ However, while the OR is a learning environment characterized by a large variety and amount of cognitive triggers, the OR experience lacks educational structure.⁷ The lack of structure can lead to much of the learning of surgical skills being based on so-called 'discovery learning', a self-guided learning approach that is based on the concept that effective learning requires mental efforts from the learner.⁸ Although discovery learning has been supported by psychologists that believe learning requires active involvement from the learner, findings within the last decennia suggest that pure discovery learning is associated with several problems relevant to surgical training, such as an overload of new information, misinterpretation of information leading to wrong constructions and a high learning inefficiency.⁸⁻¹⁰

A number of studies have demonstrated that surgical education is increasingly inhibited by environmental factors that manifest themselves on different organisational levels such as work hour restrictions¹¹⁻¹⁴, an increased focus towards patient outcomes of the public and government^{15,16} and more complex surgical cases.^{17,18} One of the most well-known changes that has caused a revolution in the way that surgery is performed, but also led to difficulties in surgical education, is the introduction of laparoscopic surgery.

Laparoscopic surgery

Extensive research has shown that laparoscopic surgery has health advantages for patients such as decreased post-operative pain, minimal blood loss and superior cosmetic results.^{19,20} It also has important socioeconomic advantages that benefit the community as a whole such as a shorter hospital admission time and a quicker reintroduction into society.²⁰ It is therefore becoming the preferred surgical technique for an increasing number of surgical procedures. On the other hand, surgeons are confronted with a list of ergonomic challenges due to an alteration of the traditional work environment during laparoscopic surgery.^{21,22} First, the surgeon works with instruments that have a long shaft and are inserted through holes in the abdominal wall. As a consequence, there is an inversion and scaling of the movements of the instruments inside the abdomen, a decreased degrees of freedom of movement and diminished tactile feedback from the operative field. Second, the surgeon derives visual input from a screen in the operating theatre that displays images from an angled camera inside the abdomen, also called laparoscope. This leads to a diversion of the viewing perspective of the surgeon from the work field. Third, the surgeon has to translate the 2D images of the instruments and the intra-abdominal structures on the monitor to a mental 3D representation in order to perform surgery. Fourth, to optimize the viewing perspective, the assisting surgeon or scrub nurse can change the camera angle inside the abdomen according to the preference of the operating surgeon. Changing the camera angle in respect to the work field leads to a deviation of the viewing axis of the laparoscope from the work axis of the surgeon, thereby increasing the difficulty of accurately moving the instruments inside the abdomen.

All of these ergonomic challenges have a negative effect on the learning curve for laparoscopic procedures.²³ So, although there are some well-established advantages to laparoscopic surgery, laparoscopic procedures are more difficult to learn and teach than open procedures. As a consequence, the introduction of laparoscopic surgery has led to debates about aptitude, training and assessment in the surgical scientific community.

Teaching laparoscopic surgery

The potential of the trainee

Evidence that supports the use of aptitude assessment can be found in research findings of occupations that place high technical demands on their workforce. In aviation, aptitude testing has been used for a long time to measure flying aptitude. Aptitude testing started during World War II, when the loss of financial resources, due to the high number of applicants not passing training criteria stimulated research about the relationship between psychometrics and flying performance.²⁴ The resulting Army Air Force Qualifying Examination included a combination of psychometric tests. The introduction of aptitude assessment was successful as it decreased the financial expense of training by reducing the number of candidates needed to obtain 100 pilot training graduates from 397 to 155.²⁴ Later research has shown that in a considerable portion of commercial and non-commercial aviation accidents errors in visual spatial perception play an important role.^{25,26} Moreover, the cumulative flying experience of pilots does not appear to influence the risk of spatial disorientation during flight.^{27,28} It is therefore believed that flying experience does not fully compensate for a low pilot aptitude.

In North American dental education, aptitude tests were introduced in the 1950s to assist dental schools in the selection of dental students.²⁹ The initial selection assignment was the Chalk Carving test in which an applicant is instructed to carve a diagram of a geometrical design into a piece of chalk. Due to logistical issues this test was replaced by a paper and pencil visual-spatial ability test in 1972.²⁹ Although in dental education the dropout rate has not been as dramatic as in the army air force, a number of study findings support the use of aptitude testing, as they have shown that the visual-spatial ability test scores can predict achievement in different levels of dental education.^{29,30}

Although technical aptitude has long been considered irrelevant or, at least, far less important than hard work in the field of surgery, laparoscopic surgery requires different skills than open surgery. It seems that the majority of surgical trainees overcome the visual-spatial and psychomotor difficulties of laparoscopic surgery during surgical training. Yet, recent research has raised concerns about individual differences before, during and after training that might be partly dependent on differences in aptitude.³¹⁻³³

Skills lab training

Nowadays, training of laparoscopic skills most often commences in a simulated minimal invasive surgery (MIS) environment. This environment involves a virtual reality (VR) simulator or a video trainer that creates a world in which the learners can safely adopt their sensor-motor system to the challenges of the MIS work environment. In VR simulators, the interface is connected to joysticks that exhibit the same ergonomic properties as laparoscopic instruments to enable manipulation and mobilisation of objects in the VR environment.³⁴ A video trainer consists of a box with holes through which a laparoscopic camera and two real laparoscopic instruments are inserted.³⁵ The images of the camera are presented on a monitor. The inside of the box functions as a playground that can only be touched with the tip of the laparoscopic instruments. Thus, in contrast to the VR simulator, the video trainer provides some tactile feedback that can be used as cognitive input.

It is generally accepted that simulator training in laparoscopic surgery leads to higher baseline abilities at the time of actual surgery on patients due to a higher degree of psychomotor adaptation to the ergonomic challenges imposed by the limitations of the work environment.^{36,37} However, the simulator computed metrics used as measurements of improvement and proficiency criteria are often chosen by testing all psychomotor metrics of a simulator in a group of novices and a group of expert laparoscopic surgeons and selecting those that show a statistical difference between the two groups.³⁸ These metrics of psychomotor skills obviously have their limitations in measuring surgical skills, leading many authors to question the validity of the training content of simulators.^{34,39} The criticism on this deficit in the content validity of simulators has been persistent and appears to remain relevant in the evaluation of more sophisticated simulators that mimic full laparoscopic

procedures.⁴⁰ As a consequence, the more complex levels of behaviour that exceed the basic sensor-motor patterns, still have to be learned through experience on patients in the OR, animal tissue or on human cadavers, all of which are costly, time absorbing and subject to medico-legal and ethical concerns. To increase the proportion of learning that can be achieved in in vitro training, the full range of intra-operative surgical skills that distinguish competent surgeons from novices during in vivo laparoscopy have to be identified.

OR training

Whether it be a Fundamentals of Laparoscopic Surgery training program on a video trainer, training on a VR simulator or training on human cadavers, preparatory skills training is followed by supervised OR training with real patients. During supervised surgical training in the OR, supervising surgeons aim to find a balance between creating the optimal learning experience and guarding the patient safety during the operation. The success of the dyadic relationship is determined by a variety of factors.

First, whether teachers teach and assess identically can hamper the learning process as it can be confusing and frustrating for a trainee to be trained by different clinical supervisors who differ in their opinion about important aspects of the operation. The differences in opinion can impede the achievement of competence, because each teacher will teach and assess in a different way. Uniform teaching and assessment criteria based on a consensus among involved supervising clinicians is therefore important for the quality of surgical training. However, training criteria based on consensus are currently absent in Dutch laparoscopic surgery training programs.⁴¹

Second, as surgeons grow in their expertise, they tend to unconsciously perform the small tasks required to achieve surgical treatment of a patient in the OR. In order to transfer the necessary skills for safe and skilful surgery to an aspiring surgeon, the teaching surgeon has to step back from unconsciously competent to consciously competent. This can be compared to teaching someone how to drive. When teaching an inexperienced driver, an experienced driver is forced to become conscious of the individual smaller steps of driving a car and has to break down the complex behaviour into small digestible actions.^{42,43} As the cognitive steps are clearly formulated, novices will be able to more easily acquire the necessary skills to become competent in driving a car on his/her own. The same principle applies to learning a surgical procedure. Untangling the complex behaviour during surgical procedures into small cognitive steps is necessary for creating an effective teaching curriculum in surgery.

Third, the surgeon has the role of clinical teacher and patient safety protector in the OR, a dual responsibility or double bind that imposes a dilemma for the surgeon as a response to one of the two responsibilities can have negative implications for the other. Supervising surgeons utilize different methods to attain an optimal balance between patient safety and education during a mentor-apprentice training model. The instructional design is characterized by the teaching surgeon positioning himself as the 'assistant surgeon' while exercising safety control management by giving verbal guidance in the form of corrections or orders and physical guidance by temporarily taking over the instruments.⁴⁴ By studying this teaching model we can understand how in many studies, the complication rate of teaching cases do not differ from cases operated on by expert surgeons.⁴⁵⁻⁴⁸ Finding the right balance between patient safety and education is a complex decision making process and is essential for the success in the teaching of surgical skills.⁴⁴

Assessment of surgical skills

The Dutch Health Agency has observed unusual complications in patients operated on during the introduction of laparoscopic surgery.⁴⁹ The government and public have therefore urged for effective training and objective assessment of laparoscopic skills during specialist training programs.⁴⁹ The majority of current surgeons were trained in a training model wherein a master surgeon decides, based on his/her own perception of the necessary skills and knowledge for surgery, whether a trainee showed sufficient improvement during surgical training. For this reason, more objective assessment methods have been developed in the last decennia.⁵⁰⁻⁵² These surgical assessment methods, such as the Objective Surgical Assessment of Technical Skills (OSATS)⁵³ and Global Operative Assessment of Laparoscopic Skills (GOALS)⁵⁴, force clinical supervisors to quantify the quality of the observed skills on a specific set of domains relevant to the development of surgical competence. The OSATS has become an integral part of assessment of surgical skills during specialist training programs in the Netherlands.⁵⁵ It can be used to monitor progression during a training program and identify strengths and weaknesses in trainees. However, it cannot be used for uniform step-wise procedure specific teaching and assessment, which is, as described above, essential for the learning process of surgical trainees. Also, the OSATS is not robust with the principles of safety control management exerted by the supervising surgeon during teaching in the OR as described previously. The last and most important disadvantage, a disadvantage that seems to be associated with all global rating scales, is the insufficient inter-rater reliability of the OSATS for high stakes examinations, such as certifications for independent practice of surgical treatment.⁵⁰

Thesis objective

The aim of this thesis is to optimize current selection, training and assessment methods in laparoscopic surgery.

Outline of thesis

This thesis is divided into 3 parts.

Part I: Aptitude

Chapter 2 describes the different forms of aptitude tests and their power in predicting laparoscopic skills in novice and in advanced surgeons. The results of the numerous studies that investigated visual-spatial ability, psychomotor ability, perceptual ability and simulator-based assessment of aptitude are aggregated and summary correlations with laparoscopic performance of groups of participants are calculated for each form of aptitude measurement to estimate the variability in laparoscopic skills that can be accounted for by aptitude and to investigate which factors influence the correlation between aptitude assessment results and laparoscopic skills.

Part II: Training

Chapter 3 evaluates a new method to identify on-the-job challenges in the operating room. The Pareto principle, also known as the '80-20 rule', is used to evaluate the verbal corrections given by surgical supervisors during training in performing a laparoscopic cholecystectomy. By analyzing the aimed corrections in behaviour, the most prevalent novice behaviours in the OR are identified for the laparoscopic cholecystectomy. The discussion in this chapter also focuses on potential methods to train the identified behaviours outside the OR.

Chapter 4 focuses on the two operation setups for the laparoscopic cholecystectomy: the French and the American position. There is a lack of evidence that one should be preferred above the other in the training of trainees. We aimed to identify an ergonomic advantage for one of the two operation setups by comparing the posture of surgeons operating in the French and the American position.

Chapter 5 focuses on developing a set of fundamental procedural steps for the laparoscopic cholecystectomy and the laparoscopic appendectomy with the help of an expert panel of abdominal surgeons.

Part III: Assessment

Chapter 6 describes important aspects of study design for estimating the inter-rater reliability of surgical skills assessment. It also addresses the interpretation and quality evaluation of inter-rater reliability calculations.

Chapter 7 describes an evaluation of the validity and reliability of an alternative assessment method to the currently used OSATS, the Global Operative Assessment of Laparoscopic Skills (GOALS).

Chapter 8 focuses on the development of a new method for the evaluation of procedural specific laparoscopic skills, independence-scaled procedural assessment, and compares this method with the OSATS and GOALS in terms of validity, reliability and support for implementation for procedural assessment.

Chapter 9 provides a general discussion on the content of this thesis and topics for future research projects.

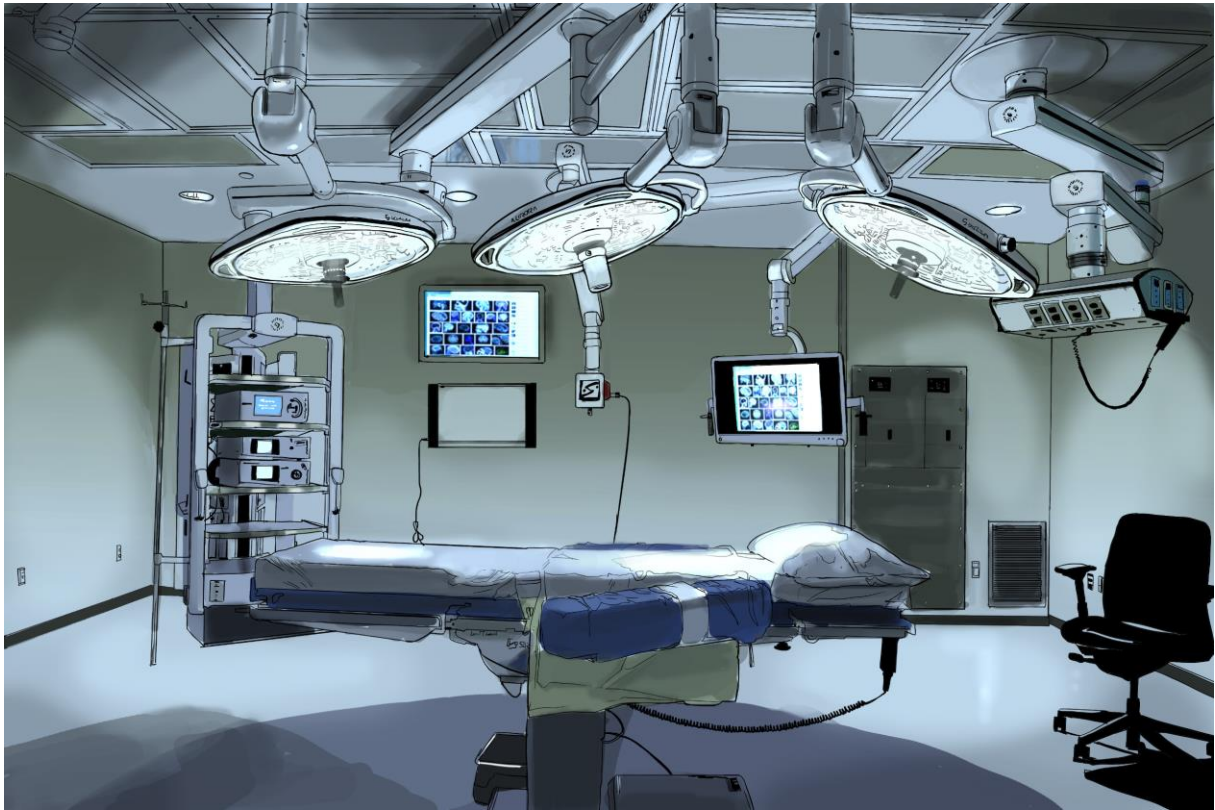
References

- 1 Pelissier C. The anthropology of teaching and learning. *Annu Rev Anthropol.* 1991;20:75-95.
- 2 Ebbinghaus H. Memory: A contribution to experimental psychology. No. 3. University Microfilms, 1913.
- 3 Ritter FE, Baxter GD, Kim JW, Srinivasmurthy S. Learning and Retention. 2012;125–142.
- 4 Speelman C, Kirsner K. Beyond the Learning Curve: The construction of mind. Oxford University Press on Demand, 2005.
- 5 Reznick RK, MacRae H. Teaching surgical skills--changes in the wind. *N Engl J Med.* 2006;355(25):2664–2669.
- 6 Lingard L, Reznick R, Espin S, Regehr G, DeVito I. Team communications in the operating room: talk patterns, sites of tension, and implications for novices. *Acad Med.* 2002;77(3):232–237.
- 7 Roberts NK, Brenner MJ, Williams RG, Kim MJ, Dunnington GL. Capturing the teachable moment: a grounded theory study of verbal teaching interactions in the operating room. *Surgery.* 2012;151(5):643–650.
- 8 Mayer RE. Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *Am Psychol.* 2004;59(1):14–19.
- 9 Aldrich NJ, Alfieri L, Brooks P, Tenenbaum HR. Does discovery-based instruction enhance learning? A meta-analysis. *J Educ Psychol.* 2007;103(1):1–18.
- 10 Kirschner PA, Sweller J, Clark RE. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ Psychol.* 2006;41(2):75-86.
- 11 Teman NR, Gauger PG, Mullan PB, Tarpley JL, Minter RM. Entrustment of General Surgery Residents in the Operating Room: Factors Contributing to Provision of Resident Autonomy. *J Am Coll Surg.* 2014;219(4):778–787.
- 12 Vollmer CM, Newman LR, Huang G, Irish J. Perspectives on intraoperative teaching: divergence and convergence between learner and teacher. *J Surg Educ.* 2011;68(6):485-494.
- 13 Ahmed N, Devitt KS, Keshet I, Spicer J, Imrie K, Feldman L, et al. A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. *Ann Surg.* 2014;259(6):1041–1053.
- 14 Hutter MM, Kellogg KC, Ferguson CM, Abbott WM, Warshaw AL. The impact of the 80-hour resident workweek on surgical residents and attending surgeons. *Ann Surg.* 2006;243(6):864–871.
- 15 Vikis EA, Mihalynuk T V, Pratt DD, Sidhu RS. Teaching and learning in the operating room is a two-way street: resident perceptions. *Am J Surg.* 2008;195(5):594–598.
- 16 Torbeck L, Wilson A, Choi J, Dunnington GL. Identification of behaviors and techniques for promoting autonomy in the operating room. *Surgery.* 2015;158(4):1102–1112.
- 17 Torbeck L, Williams RG, Choi J, Schmitz CC, Chipman JG, Dunnington GL. How much guidance is given in the operating room? Factors influencing faculty self-reports, resident perceptions, and faculty/resident agreement. *Surgery.* 2014;156(4):797–805.
- 18 Meyerson SL, Teitelbaum EN, George BC, Schuller MC, DaRosa DA, Fryer JP. Defining the autonomy gap: when expectations do not meet reality in the operating room. *J Surg Educ.* 2014;71(6):64–72.
- 19 Hendolin HI, Pääkönen ME, Alhava EM, Tarvainen R, Kempainen T, Lahtinen P. Laparoscopic or open cholecystectomy: a prospective randomised trial to compare postoperative pain, pulmonary function, and stress response. *Eur J Surg.* 2000;166(5):394-399.
- 20 Berggren U, Gordh T, Grama D, Haglund U, Rastad J, Arvidsson D. Laparoscopic versus open cholecystectomy: hospitalization, sick leave, analgesia and trauma responses. *Br J Surg.* 1994;81(9):1362-1365.
- 21 van Det MJ, Meijerink WJHJ, Hoff C, Totté ER, Pierie JPEN. Optimal ergonomics for laparoscopic surgery in minimally invasive surgery suites: a review and guidelines. *Surg Endosc.*

- 2009;23(6):1279–1285.
- 22 Hemal AK, Srinivas M, Charles AR. Ergonomic Problems Associated with Laparoscopy. *J Endourol.* 2001;15(5):499-503.
- 23 Berguer R, Smith WD, Chung YH. Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. *Surg Endosc.* 2001;15(10):1204–1207.
- 24 Griffin GR, Koonce JM. Review of psychomotor skills in pilot selection research of the U.S. military services. *Int J Aviat Psychol.* 1996;6(2):125–147.
- 25 Regan D. Spatial orientation in aviation: visual contributions. *J Vestib Res Equilib Orientat.* 1994;5(6):455–471.
- 26 Collins DL, Harrison G. Spatial disorientation episodes among F-15C pilots during Operation Desert Storm. *J Vestib Res Equilib Orientat.* 1994;5(6):405–410.
- 27 Cheung B, Money K, Wright H, Bateman W. Spatial disorientation-implicated accidents in Canadian forces, 1982-92. *Aviat Space Environ Med.* 1995;66(6):579–585.
- 28 Navathe PD, Singh B. Prevalence of spatial disorientation in Indian Air Force aircrew. *Aviat Space Environ Med.* Aerospace Medical Assn. 1994.
- 29 Ranney RR, Wilson MB, Bennett RB. Evaluation of applicants to predoctoral dental education programs: review of the literature. *J Dent Educ.* 2005;69(10):1095–1106.
- 30 Hegarty M, Keehner M, Khooshabeh P, Montello DR. How spatial abilities enhance, and are enhanced by, dental education. *Learn Individ Differ.* 2009;19(1):61-70.
- 31 Bosker R, Groen H, Hoff C, Totte E, Ploeg R, Pierie J-PP. Early learning effect of residents for laparoscopic sigmoid resection. *J Surg Educ.* 2013;70(2):200–205.
- 32 Schijven MP, Jakimowicz J. The learning curve on the Xitact {LS} 500 laparoscopy simulator: profiles of performance. *Surg Endosc.* 2004;18(1):121–127.
- 33 Grantcharov TP, Peter F-J. Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am J Surg.* 2009;197(4):447–449.
- 34 Munro MG. Surgical Simulation: Where Have We Come From? Where Are We Now? Where Are We Going? *J Minim Invasive Gynecol.* 2012;19(3):272-283.
- 35 Vassiliou MC, Dunkin BJ, Marks JM, Fried GM. FLS and FES: Comprehensive models of training and assessment. *Surg Clin North Am.* 2010;90(3):535-558.
- 36 Gurusamy K, Aggarwal R, Palanivelu L, Davidson BR. Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *Br J Surg.* 2008;95(9):1088–1097.
- 37 Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA, *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.
- 38 Korndorffer JR, Kasten SJ, Downing SM. A call for the utilization of consensus standards in the surgical education literature. *Am J Surg.* 2010;199(1):99-104.
- 39 Schout BMA, Hendriks AJM, Scheele F, Bemelmans BLH, Scherpbier AJJA. Validation and implementation of surgical simulators: a critical review of present, past, and future. *Surg Endosc.* 2009;24(3):536–546.
- 40 Bruwaene S Van, Schijven MP, Miserez M. Assessment of Procedural Skills Using Virtual Simulation Remains a Challenge. *J Surg Educ.* 2014;71(5):654-661.
- 41 Pierie J-P. De patiëntveiligheid kan worden verhoogd met structurele training in laparoscopische chirurgie. *Tijdschr voor Med Onderwijs.* 2009;28(5):196–200.
- 42 Lee FJ, Anderson JR. Does Learning a Complex Task Have to Be Complex?: A Study in Learning Decomposition. *Cogn Psychol.* 2001;42(3):267-316.
- 43 Anderson JR, Bothell D, Fincham JM, Anderson AR, Poole B, Qin Y. Brain Regions Engaged by Part- and Whole-task Performance in a Video Game: A Model-based Test of the Decomposition Hypothesis. *J Cogn Neurosci.* 2011;23(12):3983-3997.
- 44 St-Martin L, Patel P, Gallinger J, Moulton CA. Teaching the Slowing-down Moments of Operative Judgment. *Surg Clin North Am.* 2012;92(1):125–135.
- 45 Lim S, Parsa AT, Kim BD, Rosenow JM, Kim JYS. Impact of resident involvement in

- neurosurgery: an analysis of 8748 patients from the 2011 American College of Surgeons National Surgical Quality Improvement Program database. *J Neurosurg.* 2015;122(4):962-970.
- 46 Kiran RP, Ahmed Ali U, Coffey JC, Vogel JD, Pokala N, Fazio VW. Impact of resident participation in surgical operations on postoperative outcomes: National Surgical Quality Improvement Program. *Ann Surg.* 2012;256(3):469-475.
- 47 Scarborough JE, Bennett KM, Pappas TN. Defining the impact of resident participation on outcomes after appendectomy. *Ann Surg.* 2012;255(3):577-582.
- 48 Igwe E, Hernandez E, Rose S, Uppal S. Resident participation in laparoscopic hysterectomy: Impact of trainee involvement on operative times and surgical outcomes. *Am J Obstet Gynecol.* 2014;211(5):484-491.
- 49 Stassen LPS, Bemelman WA, Meijerink J. Risks of minimally invasive surgery underestimated: a report of the Dutch Health Care Inspectorate. *Surg Endosc.* 2010;24(3):495-498.
- 50 Hove PD Van, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg.* 2010;97(7):972-987.
- 51 Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. *Med Educ.* 2013; 47(7): 650-673.
- 52 Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ.* 2003;327(7422):1032-1037.
- 53 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.
- 54 Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, *et al.* A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190(1):107-113.
- 55 Opleidingsplan Heelkunde KNMG -SCHERP (Structuur Curriculum Heelkunde voor Reflectieve Professionals). [retrieved on 12-6-2016 from <http://www.knmg.nl>].

PART I: APTITUDE



Chapter 2

**THE PREDICTIVE VALUE OF APTITUDE
ASSESSMENT IN LAPAROSCOPIC
SURGERY: A META-ANALYSIS**

*Kelvin H. Kramp, Marc J. van Det, Nic J.G.M. Veeger, Christiaan Hoff, Henk O. Ten Cate
Hoedemaker, Jean-Pierre E.N. Pierie*

Based on Medical Education 2016;50(4):409-427

Abstract

Background: Current assessment methods of candidates for medical specializations that involve laparoscopic skills suffer from a lack of tools to assess the ability to work in a minimal invasive surgery environment. A meta-analysis was conducted to investigate whether aptitude assessment can be used to predict the variability in the acquisition and performance level of laparoscopic skills.

Method: PubMed, PsychInfo, and Google Scholar were searched up to November 2014 for published and unpublished studies that reported to measure a form of aptitude and laparoscopic skills. The quality of studies was assessed with QUADAS-2. The summary correlations were calculated with the random effects model.

Results: Thirty-four studies were eligible for inclusion of which six studies used an operating room performance measurement. Laparoscopic skills correlated significantly with visual-spatial ability (0.32 [95%CI 0.25 – 0.39]; $p < 0.001$), perceptual ability 0.31 ([95%CI 0.22 – 0.39]; $p < 0.001$), psychomotor ability (0.26 [95%CI 0.10 – 0.40]; $p = 0.003$) and simulator-based assessment of aptitude (0.64 [95%CI 0.52 – 0.73]; $p < 0.001$). Three-dimensional dynamic visual-spatial ability showed a significantly higher correlation than intrinsic static visual-spatial ability ($p = 0.024$).

Conclusions: In general, aptitude assessments are associated with laparoscopic skill level. Simulator-based assessment of aptitude appears to have the potential to act as a job sample by enabling the assessment of all forms of aptitude at once. A 'laparoscopy aptitude test' seems to be a valuable additional tool in the assessment of candidates for medical specializations that require laparoscopic skills.

Trial registration: PROSPERO registration: CRD42014015647

Introduction

Assessment of candidates for training in a medical discipline is a critical component of selection procedures in medical education. However, while surgical techniques are becoming increasingly difficult to master, no scientific methods are currently used to evaluate the potential to acquire these surgical skills. This is especially worrisome in the field of laparoscopic surgery, which is becoming the mainstay for an increasing list of procedures in abdominal surgery, gynecology and urology. During laparoscopic surgery, there is no direct visualization of the operative field or direct contact with intra-abdominal organs. As a consequence, a different set of skills are required compared to conventional surgery. The majority of trainees overcome the ergonomic difficulties associated with laparoscopic surgery during laparoscopic skills training, but research has raised concerns about large individual differences during and after training that might be dependent on aptitude. For instance, the studies of Schijven et al. and Grantcharov et al. both could distinguish 4 groups that showed different improvement patterns during laparoscopic skills training: 1) proficiency level from the beginning, 2) achieving proficiency level through training, 3) improvement without reaching proficiency level and 4) no improvement.^{1,2} A third study by Bosker et al. 2013 showed that one of the thirteen participants (7.7%) seemed to have problems learning to perform a laparoscopic sigmoid resection while there were no factors that could have caused a higher difficulty of the performed procedures.³

There is growing evidence in the literature that some of these differences in ability to learn and perform laparoscopic surgery can partly be explained by aptitude. Aptitude for minimally invasive surgery (MIS) can be divided into 3 abilities that are generally accepted to be of innate nature: visual-spatial ability, perceptual ability and psychomotor ability.⁴ Visual-spatial ability refers to the ability to mentally visualize and/or manipulate objects, perceptual ability refers to the ability to interpret 2D representations of 3D objects and psychomotor ability comprises motor movements like eye-hand coordination, bimanual dexterity, reaction time, etc. The question of whether testing these abilities could be used in the selection of trainees is currently a topic of vehement debate.⁵⁻⁹ While there is evidence that the ability to learn and perform laparoscopic surgery can be assessed with measurements of these aptitudes, at the same time, there have been reports that contradict such a correlation.^{1,10-13} To date, reviews aimed to reach a univocal conclusion on this topic were mainly descriptive in nature, lacked a quantitative analysis or investigated a broad spectrum of surgical skills.^{4,8,14} A meta-analysis was conducted to:

1. Evaluate whether aptitude assessments can be used to predict the ability to acquire and perform laparoscopic skills.
2. Quantify how much can be predicted by aptitude assessment.
3. Obtain insight in the factors that influence the strength of this relationship.

Method

Search Strategy

A systematic literature search in PubMed, PsychInfo and Google Scholar was conducted in November 2014 to find studies that measured laparoscopic skills and aptitude (Figure 1). The query used to identify the available literature in PubMed was ‘((Space Perception[MeSH]) OR (Visual Perception[MeSH]) OR (Psychomotor Performance[MeSH]) OR (Aptitude[Mesh])) AND (Laparoscopy[MeSH])’. The query used in PsychInfo was ‘Laparoscopy OR Laparoscopic’ with the limits age > 18 years and human studies. The queries used in Google Scholar were the word ‘laparoscopy’ combined with the exact phrase ‘visual-spatial ability’ and the word ‘laparoscopy’ combined with the exact phrase ‘psychomotor ability’. To identify unpublished literature, dissertation and thesis databases and conference abstract books were hand searched on the keywords ‘Visual-spatial’, ‘Spatial Ability’, ‘Visual Perception’, ‘Spatial Perception’, ‘Psychomotor’ and ‘Aptitude’ for additional relevant titles. Finally, the reference lists of the included studies were scanned for additional relevant studies and key authors’ names were used as search terms in PubMed and Google Scholar. Of the studies with relevant titles, abstracts were reviewed and only studies that assessed aptitude among subjects and measured laparoscopic skill level were considered eligible for inclusion.

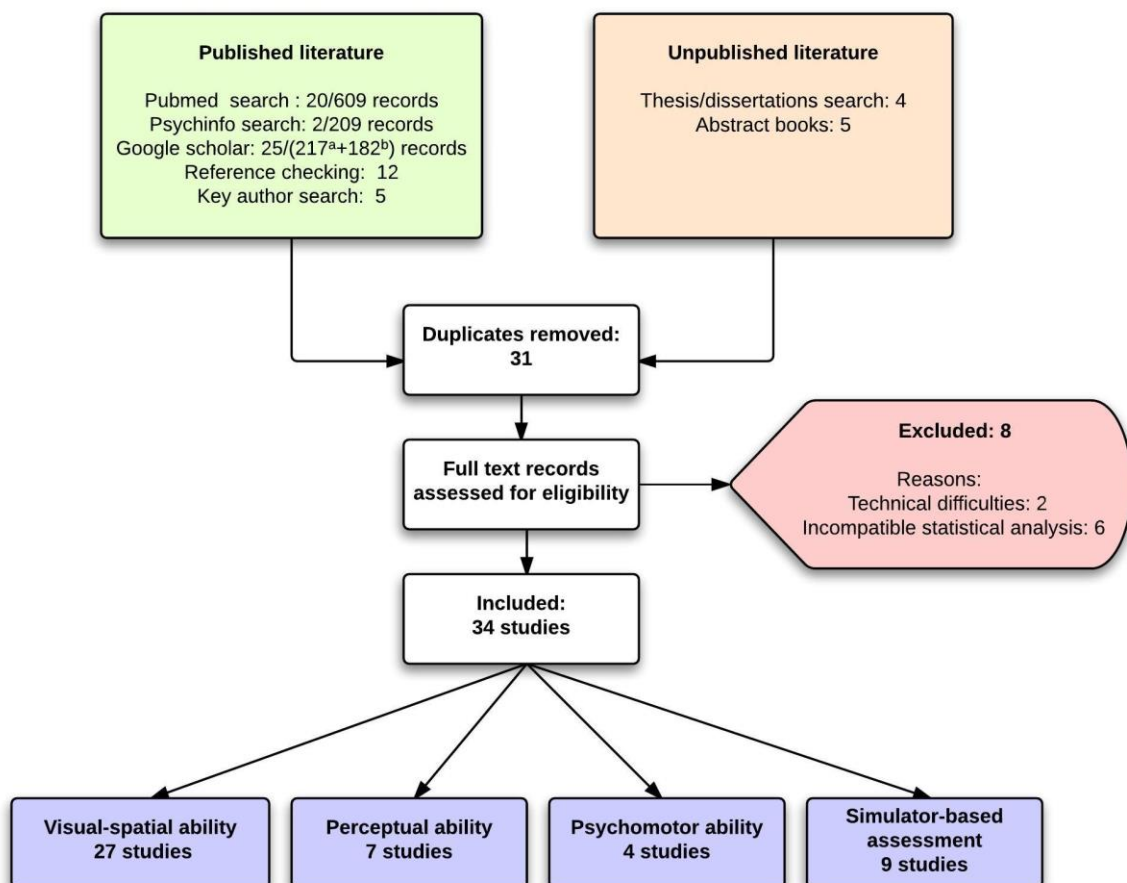


Figure 1: Flow diagram of search strategy and search results. Google scholar search: a) with exact phrase ‘visual spatial ability’ b) with exact phrase ‘psychomotor ability’. Abstract books: European Association for Endoscopic Surgery, Society of American Gastrointestinal and Endoscopic Surgeons and Association of Laparoscopic Surgeons of Great Britain and Ireland. Thesis and dissertations were found by using key author names as search terms in Google Scholar.

Data extraction and quality assessment

From the final list of studies the following information was extracted: the specific aptitude tests that were used, the reported correlations between the aptitude tests and performance level of participants, the numbers of participants, the characteristics of the participants, the measurement methods of laparoscopic skills, the parts of the learning curve in which the laparoscopic skills were measured, publication status and countries of origin.

Quality assessment of the included studies was conducted with the modified form of the QUADAS-2 (QUality Assessment of Diagnostic Accuracy Studies).¹⁵ QUADAS-2 is a quality assessment system for diagnostic studies. It contains 4 domains of assessment that should be tailored to fit the study at question: subject selection, index test, reference test and flow and timing. The first 3 mentioned domains are evaluated based on aspects related to the risk of bias and concerns related to the applicability of the study results, while the domain flow and timing is only assessed based on aspects related to risk of bias. Because the goal of this meta-analysis was to investigate the predictive validity of aptitude tests, which can be seen as an index test, the QUADAS-2 was considered as a suitable quality assessment tool.

Meta-analysis

The collected correlations were coded such that positive correlations indicate a proportional relationship and negative correlations an inverse proportional relationship between the aptitude test scores and laparoscopic performance metrics. For studies that did not report the relation between aptitude scores and laparoscopic skills with a Pearson or Spearman correlation coefficient, the reported results were converted into correlation coefficients using the formulas shown in Appendix A. The Fisher z-r transformation was used to translate the Pearson, Spearman and the converted correlations into effect sizes.

If the actual value of a non-significant correlation was not reported, lead authors were contacted for additional data. If data could not be obtained from the authors, we used two different strategies to address the missing correlation: 1) There is no relationship between the two variables or 2) the sample size is too small to achieve significance level. The first option can mathematically be considered as a correlation of 0 (Table 1). Non-significant correlations with an unreported value were therefore coded as $r = 0$ in dataset DS_0 . The second option was evaluated by substituting the maximum achievable correlation coefficient (the critical value of the Pearson correlation coefficient based on the number of participants) for the unreported non-significant correlations in dataset DS_{cv} .

Table 1: Assumptions in the dataset DS_0 and DS_{cv} .

Dataset	Assumptions	Summary estimate
DS_0	Not significant correlations with unreported size are equal to $r = 0$	Summary estimate calculated with minimum value of the correlations with unknown size.
DS_{cv}	Not significant correlations with unreported size are equal to the critical Pearson correlation coefficient	Summary estimate calculated with maximum value of the correlations with unknown size.

Some studies used multiple groups of participants with different characteristics (e.g. medical students, trainees, consultants, etc.). In these cases, the correlations for each group of participants were calculated into mean 'participant group' effect sizes. This was done by computing the mean correlation of the correlations between aptitude test outcomes and laparoscopic skill level reported in the study for a specific group of participants.

A correction was applied to the variance to compensate for the partial independence between correlation because of the commonalities in study setting in which the different correlations were measured within a participant group.¹⁶ As no correlations could be identified in the literature that could be used to correct the partial interdependence between the reported correlations, $r_x = 0.5$ was used as a compromise between the two extremes (Formula 6, Appendix A).

After estimating the participant group effect sizes and their variance, the summary correlations were calculated for the different forms of aptitude. As there was high variety

(heterogeneity) in methodology observed among the included studies, the random effects model was used to calculate the mean correlations. Heterogeneity tests (Cochrane Q) were performed to assess the variety among studies and a $p < 0.10$ was considered statistically significant. I^2 was calculated to estimate the percentage of variance that can be attributed to the variation between studies. Percentages were classified as: 25% = 'low heterogeneity', 50% = 'medium heterogeneity' and 75% = 'high heterogeneity'.¹⁷

Visual-spatial ability moderator analysis

When statistical significant heterogeneity is observed, a moderator analysis can be performed to investigate whether variation in results among studies is caused by differences in study methodology. In this meta-analysis, a moderator analysis was conducted to evaluate the different factors that could have influenced the relationship between visual-spatial ability and laparoscopic skills. Moderators were set as: 1) A recently published 2x2 classification of visual-spatial ability, 2) Measurement method of laparoscopic skills 3) Participant characteristics and 4) Components of the learning curve. The random effects model was used to calculate the summary correlation for each subgroup within the moderator analysis. A pooled τ^2 was used to estimate imprecision of subgroup summary correlations and subgroups were compared with heterogeneity Q according to the procedure described by Borenstein.¹⁶ Heterogeneity Q's with $p < 0.05$ (2-tailed) were considered statistically significant in the moderator analysis.

2x2 classification of visual-spatial ability

It is currently accepted that visual-spatial ability is not a uniform ability.^{18,19} Uttal et al. recently proposed a classification of visual-spatial ability on the basis of 2 fundamental properties of visual-spatial ability tests.¹⁸ The first distinction is whether a visual-spatial ability test utilizes intrinsic or extrinsic information. Intrinsic information contains the characteristics that define an object. Extrinsic information is information that comes from relationships between groups of objects or the relationship of an object to a framework. The second distinction is whether a visual-spatial ability test requires a static or dynamic mental process to complete. Static visual-spatial ability tests contain fixed objects, while dynamic visual-spatial ability tests require the mental visualization of a spatial change in an object or perspective. These two properties of visual-spatial ability tests can be used to classify them in 4 categories: intrinsic static, intrinsic dynamic, extrinsic static and extrinsic dynamic (Table 2). Uttal et al. suggested that a visual-spatial ability test score is an indication of the visual-spatial ability in one of the 4 categories that cannot plainly be generalized to other categories. For example, artists seem to use their visual object ability (intrinsic static) for their profession, while engineers appear to depend more on visual-spatial translational ability (intrinsic dynamic and extrinsic dynamic).^{20,21}



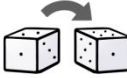

	Intrinsic (within object)	Extrinsic (between objects)
Static		
Dynamic		

Table 2: The 2x2 classification of VSA proposed by Uttah et al.: Intrinsic static, intrinsic dynamic, extrinsic static and extrinsic dynamic.

Other moderators

There were 3 other moderators used to evaluate heterogeneity between studies in visual-spatial ability: measurement method, participant characteristics and part of the learning curve in which the performance level was measured.

A typical learning curve has been documented in laparoscopic skills training on video trainers, virtual reality simulators and in the OR (Table 3).^{22,23} The learning curve starts with a baseline performance. After the first performance there are improvements with each repetition. The improvements decrease in size as experience accrues according to the learning rate. As the improvements become increasingly small with each repetition a learning plateau is reached.²⁴ The strength of the association between visual-spatial ability test scores and laparoscopic skills could differ between the learning phases. In the moderator learning curve, the correlations were therefore categorized as: 1) Baseline performance, 2) Learning rate and 3) Learning plateau.

The association between visual-spatial ability and laparoscopic skills could differ on the basis of the method used to measure skill level. To investigate whether the method of measuring performance influences the correlation with visual-spatial ability, the correlations were categorized as follows: 1) Video trainer, 2) Virtual reality simulator, 3) Laparoscopic camera navigation and 4) Laparoscopic surgery on an animal or human.

Whether the study sample represents the characteristics of the population of interest can be of key importance in the validation of prognostic tools. In the moderator participant characteristics, the correlations were therefore categorized as: 1) Non-medical students, 2) Medical students, 3) Novice trainees and 4) Trainees who had received training in laparoscopy and consultants.

Table 3: Definition of baseline performance, learning rate and learning plateau.

Learning phase	Definition
Baseline performance	Measurement when exposed to a task (e.g. first performance or average of repetition 1 to 3) of which the content has not been repeatedly performed on a VRS, VT or in the OR.
Learning rate	Rate of improvement defined by the measured slope or exponential of the learning curve.
Learning plateau	Measurement when skill level shows no evidence of significant improvement when a task is repeatedly performed.

Publication bias

To visually assess the sample of studies for publication bias, a funnel plot was created with 95% pseudo confidence intervals. Quantitative assessment of publication bias was performed with the Egger and Begg tests with p-values < 0.05 (2-tailed) considered significant.^{25,26}

Results

Search results

The results of the literature search are shown in figure 1. Eight studies were excluded because of a methodology or statistical analysis incompatible with the research question of the meta-analysis or because of technical difficulties during aptitude testing.²⁷⁻³⁴ Some studies evaluated the correlation between a simulator-based assessment and a subsequent performance on a simulator or performance in the OR. These studies were analyzed separately from the other aptitude measurements. In total 34 studies were eligible on the basis of the inclusion criteria and could be used for further analysis of which 6 studies included an OR performance measurement. A limited series of studies reported cut-off scores for the classification of candidates and their corresponding sensitivity and specificity (Table 4).^{1,35,36}

Table 4: Reported cut-off scores for various VSA and PMA tests and for simulator baseline performance (BL) and training outcome.

Aptitude test	Test	Classification	Cut-off score	Sensitivity	Specificity	Author
VSA	Rey Figure	Slow learners	<21.5	60	80	Stefanidis ³⁵
	Map Planning	Slow learners	<16	60	100	Stefanidis ³⁵
	Space Relations	Likely to encounter problems during training	<25	NA	NA	Schijven ¹
	Space Relations	Unlikely to encounter problems during training	>45	NA	NA	Schijven ¹
PMA	Finger Tapping	Slow learners	<61	100	47	Stefanidis ³⁵
	Grooved Pegboard	Slow learners	>54	80	47	Stefanidis ³⁵
SPM	VT BL	Slow learners	>473	60	87	Stefanidis ³⁵
	LCN BL	Slow learners	>238	60	60	Stefanidis ³⁵
	FLS training	GOALS OR performance \geq level of experienced surgeon	>70	91	86	McCluney ³⁶

VT = video trainer , LCN = laparoscopic camera navigation, FLS = Fundamentals of Laparoscopic Surgery, BL = baseline performance, VT BL performance measurement consist of the average score of 3 repetitions of 5 Southwestern stations.

LCN BL performance measurement consists of the average score of 3 repetitions of 2 different LCN tasks on Tulane Camera Navigation Simulator.

GOALS (Global Assessment of Laparoscopic Skills) is a subjective assessment that can be used to evaluate 5 measures of laparoscopic skills: depth perception, bimanual dexterity, efficiency, tissue handling and autonomy.

Quality assessment

The quality of the 34 included studies is shown in Appendix C. Seven studies reported correlations that were not at risk of bias and 6 studies used an OR performance measurement and therefore did not raise concerns regarding applicability.^{13,36-47}

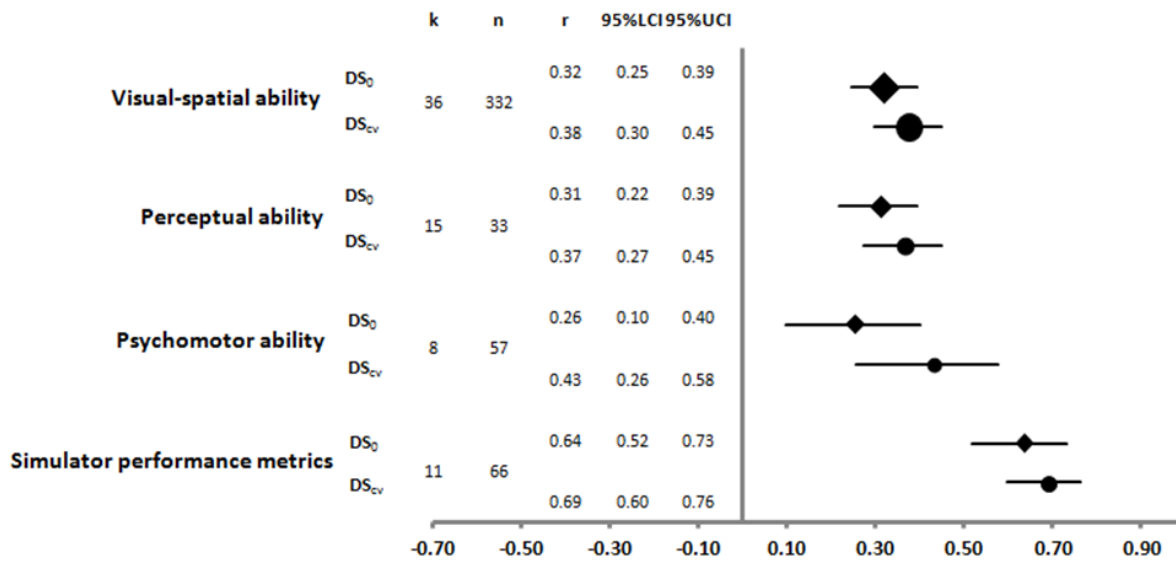


Figure 2: Forest plot of the summary correlations of the different aptitude measurements with their 95% confidence interval. DS₀: Dataset with not reported not significant correlations coded as 0. DS_{cv}: Dataset with not reported not significant correlations coded as the critical Pearson correlation coefficient at $\alpha_2 = 0.05$ and $dF = N - 2$. k = number of participant groups, n = number of correlations, r = summary correlation, LCI = lower confidence interval, UCI = upper confidence interval.

Visual-spatial ability and laparoscopic skills

Twenty-seven studies containing 36 participant groups were included in the analysis of visual-spatial ability (Table 5). An overview of the encountered aptitude tests for visual-spatial ability is shown in Appendix C. In 5 groups of participants laparoscopic skills were measured in the OR.^{13,47} The mean correlation in DS₀ (dataset with the unreported non-significant correlations coded as 0) was 0.32 (95%CI [0.25-0.39], $p < 0.001$), Q was statistically significant (74.55, $p < 0.001$) and I^2 was 53%, indicating above moderate heterogeneity (Figure 2). When the inclusion of studies was limited to studies that used an OR performance to measure laparoscopic skills, the mean correlation in DS₀ increased to 0.50 (95%CI [0.07-0.77], $p = 0.024$).

Table 5: Overview of studies measuring the predictive power of visual-spatial ability.

	Author	Year	N	Level of training	VSA	MM	BL	LR	LP	Sign	Correlation	PS	Remarks
1	Risucci ⁶⁶	2000	39	20 beginning PGY 1 19 PGY 3-5 (Np > 30)	1,2b	VT	NA	NA	NA	3/12	0.41 - 0.71	p	
2	Eyal ⁴¹	2001	27	Undergraduates	2a,2 b,4	LCN	NA	NA	NA	4/4	0.39 - 0.58	p	
3	Risucci ⁶⁷	2001	94	23 PGY >3 71 attending surgeons	1,2b	VT	+	NA	+	10/1 2	0.21 - 0.51	p	Partial correlation corrected for acquired knowledge during course also significant
4	Haluck ⁶⁸	2002	25	No simulator experience	2a,2 b,3	LCN	NA	NA	NA	NA/3	0.30-0.39	p	
5a	Keehner ¹³	2004 a	48	Low experience (median Np = 13)	2b	OR	NA	NA	NA	1/1	0.39	p	Animals
5b	Keehner ¹³	2004 b	45	High experience (median Np = 302)	2b	OR	NA	NA	NA	0/1	0.02	p	Animals
6	Schijven ¹	2004	28	Hospital residents and final year interns	2b	VRS	NA	NA	NA	0/1	0.40	p	Kendall tau correlation
7	McClusky ¹²	2005	11	4th year medical students	2a,2 b,3	VRS	NA	NA	NA	0/3	-	p	
8	Stefanidis ³⁵	2006	20	First year surgical residents (median Np = 0)	1,2a, 2b,3	VT/V RS/L	+	NA	NA	5/30	0.44 - 0.64	p	

9a	Hedman ⁴²	2006 a	54	Medical students (N=0)	2b	CN	VRS	+	NA	NA	8/15	0.28 - 0.40	p	
9b	Hedman ⁴²	2006 b	25	Medical students (N=0)	2b	VRS	VRS	+	NA	NA	8/15	0.43 - 0.49	p	Correlation significant when corrected for BASIQ-general IQ test scores
10	Keehner ³⁷	2006	22	Non medical students	5	LCN	LCN	+	NA	+	3/4	0.19 - 0.46	p	Correlation corrected for general intelligence significant
11	Birbas ⁶⁹	2006	21	Minimal experience	5	VRS	VRS	NA	NA	+	1/NA	0.72	a	
12	Andalib ⁷⁰	2006	32	Medical and dental students (Ns = 0)	2a,2 b,3,5	VRS	VRS	NA	+	+	3/6	0.36 - 0.46	a	
13	Hassan ⁷¹	2007	16	NA	2b	VRS	VRS	NA	NA	NA	NA/NA	NA	p	Mann Whitney U test
14	Enochsson ⁷²	2008	9	Gynecological consultants	2b	VRS	VRS	NA	NA	NA	4/4	0.72 - 0.82	a	
15	Rosenthal ³⁰	2010	56	Novice (Np = 0) to expert (Np > 100)	2b	LCN	LCN	NA	NA	NA	4/3	0.28 - 0.45	p	
16	Sliwinski ⁷³	2010	7	Surgical and gynecologic trainees	1,2a, 2b	VRS	VRS	-	NA	+	2/20	0.78 - 0.88	t	
17	Kolozsvari ⁷⁴	2011	32	Medical and dental students (Ns = 0)	2a,2 b,3	VT	VT	-	-	-	0/9	-	p	Only study that measured surgical interest
18	Jungmann ⁴³	2011	40	Medical students (Ns = 0)	2b	VRS	VRS	NA	NA	NA	2/3	0.38 - 0.56	p	
19	Ahlborg ⁴⁴	2011	13	Gynecological consultants	2b	VRS	VRS	NA	NA	NA	7/15	0.57-0.64	p	
20	Schlickum ⁷⁵	2011	25	Medical students	2b	VRS	VRS	NA	NA	NA	1/1	0.45	p	
21	Luursema ¹¹	2012	23	Technical medicine students	1,2a, 5	VRS	VRS	-	NA	-	0/3	-	p	
22a	Ahlborg ³⁸	2012 a	28	Gynecological trainees (Ns = 0, Np < 10)	2b	VRS	VRS	+	NA	NA	1/2	0.40	p	
22b	Ahlborg ³⁸	2012 b	19	Gynecological trainees /w VRS training	2b	VRS	VRS	NA	NA	NA	1/2	-	p	
23a	Nugent ⁴⁵	2012 a	40	Pre-clinical medical students Y1-3	2a,2 b,3	VRS	VRS	+	NA	NA	9/12	0.34 - 0.48	d	
23b	Nugent ⁴⁵	2012 b	20	12 PGY 1 basic surgical trainees 8 PGY 2 basic surgical trainees	2a,2 b,3	VRS	VRS	NA	NA	NA	3/12	0.45 - 0.59	d	
23c	Nugent ⁴⁵	2012 c	8	Higher surgical trainees Y1-3	2a,2 b,3	VRS	VRS	NA	NA	NA	3/12	0.75 - 0.80	d	
23d	Nugent ⁴⁵	2012 d	12	Higher surgical trainees Y4-6	2a,2 b,3	VRS	VRS	NA	NA	NA	0/12	-	d	
23e	Nugent ⁴⁵	2012 e	26	Pre-clinical medical students Y 1-3	0	VRS	VRS	NA	NA	NA	3/4	0.54 - 0.94	d	
24	Nugent ⁴⁵	2012	67	General and plastic surgery trainees	0	VRS/ BM	VRS/ BM	NA	NA	NA	0/2	-	d	
25	Nugent ⁷⁶	2012	10	Surgical trainees (Nbl > 20, Nal < 5)	0	VRS	VRS	+	NA	NA	6/13	0.67 - 0.78	p	
26a	Ahlborg ⁴⁷	2013	28	Gynecological trainees (Ns = 0, Np < 10)	2b	OR	OR	+	NA	NA	0/1	0.33	p	Humans
26b	Ahlborg ⁴⁷	2013	7	Gynecological trainees no VRS training	2b	OR	OR	+	NA	NA	1/1	0.98	p	Humans
26c	Ahlborg ⁴⁷	2013	13	Gynecological trainees /w VRS training	2b	OR	OR	NA	NA	NA	0/1	0.13	p	Humans
27	Groenier ¹⁰	2014	53	Technical medicine students	1,2b	VRS	VRS	NA	NA	NA	0/1	-	p	

MM = measurement method, VT = video trainer, VRS = Virtual Reality simulator, LCN = laparoscopic camera navigation, OR = animal/human OR performance, BM = Bench Models.
BL = baseline performance, LR = learning rate, LP = learning plateau. NA = not addressed, +/- significant/not significant.

PS = publication status: a = abstract, d = dissertation, t = thesis, p = published in peer review journal.

N = number of participants, Ns = Number of performed simulator tasks, Np = Number of performed procedures, Nbl = Number of performed basic laparoscopic procedures, Nal = Number of performed advanced laparoscopic procedures.

VSA = visual-spatial ability: 0 = composite score static and dynamic, 1 = intrinsic static, 2a = Intrinsic dynamic 2D, 2b = Intrinsic dynamic 3D, 3 = extrinsic static, 4 = extrinsic dynamic and 5= composite score intrinsic dynamic and extrinsic dynamic.

PGY = post graduate year.

Correlation: min. and max. significant correlation found in the study.

Perceptual ability and laparoscopic skills

In all included studies perceptual ability was assessed with the Pictorial Surface Orientation test (PicSOr). The PicSOr was developed to measure the ability to recognize the 3D orientation of an virtual object from a 2D screen.^{12,48} Seven studies containing 15 participant groups were included in the correlation analysis of perceptual ability (Table 6). The mean correlation for perceptual ability in DS₀ was 0.31 (95%CI [0.22-0.39], p < 0.001), Q was 21.30 (p = 0.128), indicating no heterogeneity among studies (Figure 2).

Table 6: Overview of studies measuring the predictive power of perceptual ability.

	Author	Year	N	Level of training	MM	BL	LR	LP	Sign	Correlation	PS
1	Haluck ⁶⁸	2002	25	No simulator experience	LCN	NA	NA	NA	1/1	0.59	P
2a	Gallagher ⁴⁸	2003	48	Laparoscopic novices	VT	NA	NA	NA	1/1	0.50	P
2b	Gallagher ⁴⁸	2003	32	Laparoscopic novices	VT	NA	NA	NA	1/1	0.50	P
2c	Gallagher ⁴⁸	2003	34	Laparoscopic novices and experienced surgeons	VT	NA	NA	NA	1/1	0.42	P
2d	Gallagher ⁴⁸	2003	18	Experienced laparoscopic surgeons	VT	NA	NA	NA	1/1	0.54	P
3	McClusky ¹²	2005	11	4 th year medical students	VRS	NA	NA	NA	1/1	0.76	P
4	Stefanidis ³⁵	2006	20	1 st year surgical residents (median Np = 0)	VT/VRS/LCN	-	NA	NA	0/5	NS	P
5	Kolozsvari ⁷⁴	2011	32	Medical and dental students (Ns = 0)	VT	+	-	-	1/3	0.38	p
6a	Nugent ⁴⁵	2012 a	40	Pre-clinical medical students Y1-3	VRS	+	NA	NA	1/3	0.49	d
6b	Nugent ⁴⁵	2012 b	20	12 PGY 1 basic surgical trainees 8 PGY 2 basic surgical trainees	VRS	NA	NA	NA	1/3	0.52	d
6c	Nugent ⁴⁵	2012 c	8	Higher surgical trainees Y1-3	VRS	NA	NA	NA	1/3	0.80	d
6d	Nugent ⁴⁵	2012 d	12	Higher surgical trainees Y4-6	VRS	NA	NA	NA	2/3	0.70 - 0.73	d
6e	Nugent ⁴⁵	2012 e	26	Pre-clinical medical students Y1-3	VRS	NA	NA	NA	1/4	0.56	d
7	Nugent ⁴⁵	2012	67	General and plastic surgery trainees	VRS/BM	NA	NA	NA	1/2	0.31	d

MM = measurement method, VT = video trainer, VRS = Virtual Reality simulator, LCN = laparoscopic camera navigation, OR = animal/human OR performance, BM = Bench Models.

BL = baseline performance, LR = learning rate, LP = learning plateau, NA = not addressed, +/- = significant/not significant.

PS = publication status: a = abstract, d = dissertation, t = thesis, p = published in peer review journal.

N = number of participants, Ns = number of performed simulator tasks, Np = number of performed procedures,

PGY = post graduate Year.

Correlation: min. and max. significant correlation found in the study.

Psychomotor ability and laparoscopic skills

Four studies containing 8 participant groups were included in the correlation analysis of psychomotor ability (Table 7). An overview of the encountered aptitude tests for psychomotor ability is shown in Appendix D. Of the 7 different psychomotor ability tests used, only the Finger tap test and the Grooved Peg Board test showed a significant correlation with laparoscopic performance.^{35,45} The mean correlation for psychomotor ability in DS₀ was 0.26 (95%CI [0.10-0.40], p < 0.003), Q was 9.85 (p = 0.197), indicating no statistically significant heterogeneity among studies (Figure 2).

Table 7: Overview of studies measuring the predictive power of psychomotor ability.

	Author	Year	N	Level of training	PMA test	MM	BL	LR	LP	Sign	Correlation	PS
1	Schijven ¹	2004	28	Hospital residents and final year interns	GSM, CSPDT	VRS	NA	NA	NA	0/3	NS	p
2	Stefanidis ³⁵	2006	20	1 st year surgical residents	Tremor, Reaction	VT/VRS/L	-	NA	NA	3/25	0.56-0.67	p

				(median Np = 0)	time, Finger tap, Purdue PEG board, Grooved PEG board	CN						
3a	Nugent ⁴⁵	2012 a	40	Pre-clinical medical students Y 1-3	Grooved PEG board	VRS	NA	NA	NA	2/3	0.38-0.45	d
3b	Nugent ⁴⁵	2012 b	20	12 PGY 1 basic surgical trainees 8 PGY 2 basic surgical trainees	Grooved PEG board	VRS	NA	NA	NA	2/3	0.48-0.69	d
3c	Nugent ⁴⁵	2012 c	8	Higher surgical trainees Y 1- 3	Grooved PEG board	VRS	NA	NA	NA	3/3	0.75-0.78	d
3d	Nugent ⁴⁵	2012 d	12	Higher surgical trainees Y 4- 6	Grooved PEG board	VRS	NA	NA	NA	1/3	0.7	d
3e	Nugent ⁴⁵	2012 e	26	Pre-clinical medical students Y 1-3	Grooved PEG board	VRS	NA	NA	NA	0/4	NS	d
4	Nugent ⁷⁶	2012	10	Surgical trainees (Nbl > 20, Nal < 5)	Grooved PEG board	VRS	+	NA	NA	5/13	0.77-0.87	p

MM = measurement method, VT = video trainer, VRS = Virtual Reality simulator, LCN = laparoscopic camera navigation, OR = animal/human OR performance.

BL = baseline performance, LR = learning rate, LP = learning plateau. NA = not addressed, +/- = significant/not significant.

PS = publication status: a = abstract, d = dissertation, t = thesis, p = published in peer review journal.

N = number of participants, Np = number of performed procedures, Nbl = number of performed basic laparoscopic procedures, Nal = number of performed advanced laparoscopic procedures.

PGY = post graduate Year.

GSM = Gibson Spiral Maze, CSPDT = Crawford Small Parts Dexterity Tester.

Correlation: min. and max. significant correlation found in the study.

Simulator-based assessment

Nine studies containing 11 participant groups were included in the analysis of the predictive validity of simulator-based assessment of aptitude (Table 8). In 5 out of 9 studies laparoscopic skills training parameters were correlated with OR performance measurements.^{36,39,40,46,49} The mean correlation for simulated MIS performance in DS₀ was 0.64 (95%CI [0.52-0.73], p < 0.001), Q was 13.78 (p = 0.183), indicating no statistically significant heterogeneity among studies (Figure 2). When the inclusion of studies was limited to the correlation between simulator performance and a subsequent OR performance, the mean correlation in DS₀ decreased to 0.61 (95%CI [0.42-0.75], p < 0.001).

Table 8: Overview of studies measuring the predictive power of simulation-based assessment.

	Author	Year	N	Level of training	Aptitude test	MM	BL	LR	LP	Sign	Correlation	PS
1	Macmillan ⁴⁶	1999	10	Higher surgical trainees	ADEPT	OR	NA	NA	NA	3/3	0.74-0.79	p
2a	Chaudhry ⁷⁷	1999	7	Hospital staff	VRS BL	VRS	NA	NA	+	4/6	0.01-1.00	p
2b	Chaudhry ⁷⁷	1999	11	Basic surgical trainees and above	VRS BL	VRS	NA	NA	+	2/6	0.61-0.89	P
2c	Chaudhry ⁷⁷	1999	17	Medical students	VRS BL	VRS	NA	NA	+	6/6	0.56-0.98	P
3	Ahlberg ³⁹	2002	12	Medical student	VRS	OR	+	NA	NA	2/2	0.33-0.64	P
4	McClusky ¹²	2005	11	4 th year medical students	VRS	duration of training	NA	NA	NA	2/2	0.62-0.73	p
5	Stefanidis ³⁵	2006	20	1st year surgical residents (median Np = 0)	VT/VRS/LC N BL	duration of training	NA	NA	NA	4/6	0.55-0.66	p
6	McCluney ³⁶	2007	40	Surgical trainees PGY1-5, surgical fellows and consultant surgeons	FLS	OR (GOALS)	NA	NA	NA	1/1	0.77	P
7	Hogle ⁴⁰	2008	10	Surgical trainees PGY1	VRS	OR (GOALS)	NA	NA	NA	0/1	NS	p
8	Kundhal ⁴⁹	2009	10	Surgical trainees Np = 5	VRS/LCN	OR (OSATS)	+	NA	NA	19/28	0.67-0.98	p
9	Nugent ⁴⁵	2012	10	Surgical trainees (Nbl > 20, Nal < 5)	VRS basic tasks	VRS colectomy	+	NA	NA	3/6	0.77-0.92	d

MM = measurement method, VT = video trainer, VRS = Virtual Reality simulator, LCN = laparoscopic camera navigation, OR = animal/human OR performance.

BL = baseline performance, LR = learning rate, LP = learning plateau. NA = not addressed, +/- = significant/not significant.

PS = publication status: a = abstract, d = dissertation, t = thesis, p = published in peer review journal.

N = number of participants, Np = Number of performed procedures, Nbl = Number of performed basic laparoscopic procedures, Nal = number of performed advanced laparoscopic procedures.

PGY = Post graduate Year.

ADEPT = Advanced Dundee Endoscopic Psychomotor Tester, FLS = Fundamentals of Laparoscopic Surgery, PGY = Post graduate Year, OSATS = Objective Surgical Assessment of Technical Skills, GOALS = Global Operative Assessment of Laparoscopic Skills.

Correlation: min. and max. significant correlation found in the study.

Visual-spatial ability moderator analysis

2x2 classification of visual-spatial ability

There was significant heterogeneity in the summary correlation of visual-spatial ability ($p < 0.001$). A moderator analysis was performed to investigate whether the heterogeneity among studies is caused by differences in methodology. The results of the moderator analysis for visual-spatial ability are shown in table 9. In the analysis of the 2x2 classification of visual-spatial ability, two studies were excluded because they used a composite measure of static and dynamic visual-spatial ability tests. The subgroup extrinsic dynamic contained only one study.³⁷ The subgroups intrinsic dynamic and extrinsic dynamic were thus combined into the subgroup dynamic visual-spatial ability. Close inspection of this subgroup showed a subdivision of 2D visual-spatial ability tests, often of low complexity. The subgroup dynamic visual-spatial ability was consequently divided into Dynamic 2D and Dynamic 3D to create an adjusted 2x2 classification.

The subgroup intrinsic static showed no statistically significant correlation in DS_0 (dataset with the unreported non-significant correlations coded as 0) and DS_{cv} (dataset with the unreported non-significant correlations coded as maximum achievable correlation) (resp. $p = 0.069$ and $p = 0.100$) and the subgroup extrinsic static showed no significant correlation only in DS_0 ($p = 0.075$). A significant difference was observed between subgroups ($p = 0.024$). The unknown size of non-significant correlations led to a substantial difference between subgroups in DS_0 and DS_{cv} in the subgroup extrinsic static (DS_0 : $r=0.14$, (95%CI [-0.01 - 0.28]); DS_{cv} : $r=0.34$, (95%CI [0.19 - 0.48])). Consequently, only the subgroups 3D dynamic and intrinsic static were mutually compared. Comparison of the 3D dynamic and intrinsic static subgroup showed a statistically significant difference ($p = 0.024$).

Other moderators

In the moderator measurement method, participant characteristics and learning curve no significant difference was observed between subgroups (resp. $p = 0.553$, $p = 0.271$ and $p = 0.507$, table 9).

Interestingly, a significant correlation was observed in the subgroup learning plateau in the moderator learning curve and in the subgroup trained participants in the moderator participant characteristics.

Table 9: Results of the moderator analysis for visual-spatial ability.

Moderator	Subgroup	k	n	r	95%LCI	95%UCI	p _r	Q	p _Q	I ²	p _{mod}
Adjusted 2x2 classification	Intrinsic static	6	74	0.14	-0.01	0.29	0.069	4.61	0.466	0	0.024
	Extrinsic static	9	29	0.14	-0.01	0.28	0.075	6.21	0.624	0	
	Dynamic 2D	12	59	0.21	0.08	0.34	0.002	9.11	0.612	0	
	Dynamic 3D	32	151	0.33	0.26	0.39	0.000	59.91	0.001	48	
Moderator	Subgroup	k	n	r	95%LCI	95%UCI	p _r	Q	p _Q	I ²	p _{mod}
Learning curve	BL	13	69	0.23	0.10	0.36	0.002	34.16	0.001	65	0.507
	LR	2	6	0.05	-0.26	0.35	0.381	2.48	0.289	60	
	LP	7	66	0.26	0.08	0.41	0.007	3.41	0.844	0	
Moderator	Subgroup	k	n	r	95%LCI	95%UCI	p _r	Q	p _Q	I ²	p _{mod}
Measurement method	VT	5	51	0.21	0.01	0.38	0.044	5.29	0.381	24	0.553
	VRS	22	231	0.32	0.22	0.42	0.000	44.81	0.003	53	
	LCN	7	32	0.34	0.18	0.49	0.000	2.67	0.914	0	
	OR	5	5	0.40	0.17	0.59	0.002	23.93	0.000	83	
Moderator	Subgroup	k	n	r	95%LCI	95%UCI	p _r	Q	p _Q	I ²	p _{mod}
Participant characteristics	Non-medical students	3	102	0.19	-0.08	0.42	0.154	1.71	0.635	0	0.271
	Medical students	10	56	0.33	0.19	0.45	0.000	18.85	0.042	52	
	Novice trainees	10	85	0.40	0.25	0.54	0.000	20.78	0.023	57	
	Trained participants	9	66	0.21	0.04	0.37	0.021	16.49	0.057	51	

Between group variance among moderators was evaluated with heterogeneity Q in a mixed effects model. A pooled τ^2 among subgroups was used to estimate random effects model summary estimate within subgroups. $P < 0.05$ was considered statistical significant.

k = number of groups of participants, n = number of correlations in subgroup, LCI = lower confidence interval, LCU = upper confidence interval, p_r = p-value of z-score of mean correlation, Q = heterogeneity Q, p_Q = p-value of heterogeneity Q, p_{mod} = p-value of between group variance.

Measurement method: VT = video trainer, VRS = VR Simulator, LCN = laparoscopic camera navigation, OR = operating room.

Participants characteristics: Trained participants = trainees with training in laparoscopic surgery and consultant specialists.

Learning curve: BL = baseline performance, LR = learning rate, LP = learning plateau.

Publication bias

Visual inspection of the funnel plot of visual-spatial ability showed an asymmetric distribution of the participant group effect sizes in DS₀ (Figure 3a). The Begg test and Egger test were both significant (resp. p = 0.014 and p = 0.006), indicating the possibility of publication bias. The most evident outliers were the studies of Ahlborg et al. (right lower quadrant) and Nugent et al. (left upper quadrant).^{45,47}

The small participant group of Ahlborg et al. was the only group of novice trainees performing in the OR without prior simulator training. The higher task difficulty, in comparison to the more commonly used simulator tasks could have enlarged the measurable range in skill level, leading to a higher correlation. The large study of Nugent et al. was the only study that measured basic laparoscopic skills in a subgroup of highest scoring trainees after full basic surgical training. The pre-selection of highest scoring trainees might have led to a range restriction in laparoscopic ability, and as a consequence, the observation of a low correlation in this study. Thus, study methodology probably had an opposite effect on the measurable range of laparoscopic skills in the small and large participant group that were visually identified as outliers. After removal of these outliers, the funnel

plot was symmetric and the Begg test and Egger test were not significant (resp. $p = 0.075$ and $p = 0.067$). Therefore, the evidence for publication bias was probably caused by the differences in methodology of the two included studies.

In the evaluation of publication bias for perceptual ability (PicSO_r) the funnel plot showed absence of participant group effect sizes in the left lower quadrant (Figure 3b), although the Begg test and Egger test did not indicate the presence of publication bias (resp. $p = 0.171$ and $p = 0.090$). To exclude methodology as a potential cause of bias, the characteristics of the studies in the right lower quadrant of the funnel plot were inspected.²⁶ No common difference in methodology was observed in these studies. However, retrospective evaluation of the three excluded articles that addressed perceptual ability showed that these studies reported low correlations.^{29,33,34} These studies were not included in the meta-analysis because they reported technical errors during data acquisition. Therefore, exclusion criteria could have induced bias across studies examining the PicSO_r. Other causes of the asymmetrical shape of the funnel plot that could not be excluded are publication bias and coincidence.²⁶

The Begg test and Egger test were not significant for psychomotor ability (resp. $p = 0.083$ and $p = 0.086$), but the funnel plot showed a small participant group with a large effect size. No difference in methodology could be identified in this participant group (Figure 3c).⁴⁵ Exclusion of this outlier led to a smaller, but still significant mean correlation of 0.22 (95%CI [0.08-0.35]; $p = 0.004$).

The funnel plot, the Begg test and Egger test did not indicate the presence of publication bias for simulator-based assessment of aptitude (resp. $p = 0.756$ and $p = 0.408$, Figure 3d).

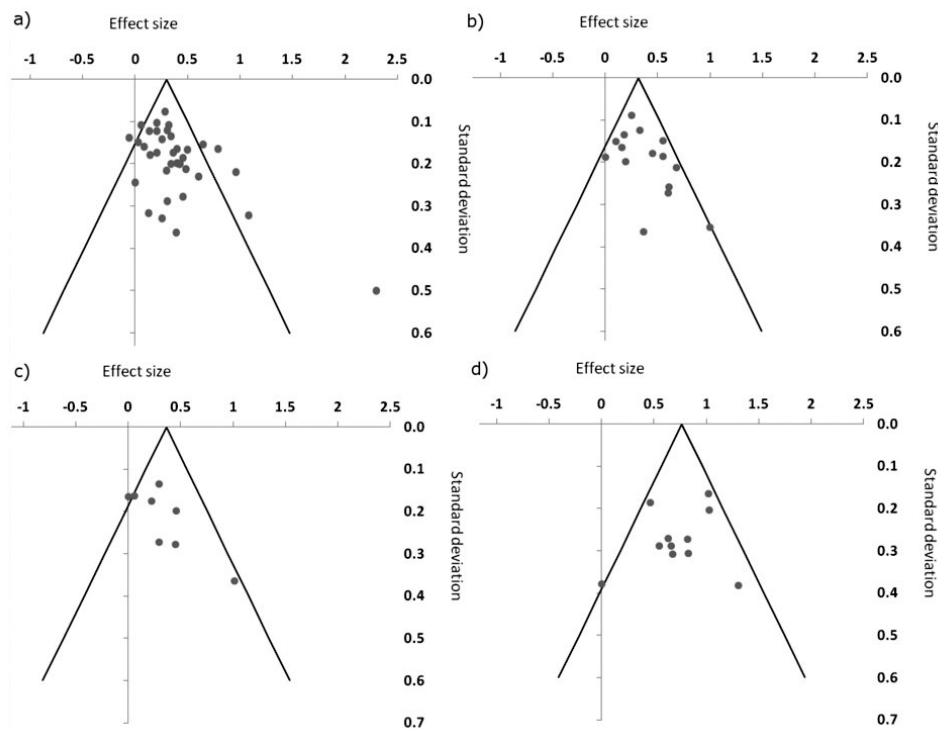


Figure 3: Funnel plot of studies measuring the correlation of laparoscopic performance with a) Visual-spatial ability, b) Perceptual ability, c) Psychomotor ability and d) Simulator-based assessment of aptitude.

Discussion

Multiple studies have shown that medical students with an interest in pursuing a surgical career display an equal variety in aptitude as medical students that are not interested in surgery.⁵⁰⁻⁵² Currently used assessment methods for medical specializations that require laparoscopy do not provide information about the potential to learn and perform laparoscopic skills to faculty members responsible for the assessment of trainees.⁵³ It has long been recognized in psychology that visual-spatial ability, perceptual ability and psychomotor ability determine performance level in technical professions to some extent.⁵⁴⁻⁵⁶ The results of this meta-analysis demonstrate that aptitude tests can be used to predict part of the individual differences in learning and performing laparoscopic skills. Aptitude test can therefore be considered as a useful adjunct to the currently used assessment methods (Figure 4). A laparoscopy aptitude test could also help students or trainees make the right career decision and/or support surgical educators in the recommendation to opt for an area of medicine that matches their talent. Persons tested with a high aptitude interested in a non-surgical career obtain a stimulus to consider pursuing a surgical career and those with a low aptitude interested in a career involving laparoscopy have the opportunity to invest their valuable time and energy in a specialty or differentiation program that better matches their talent at an earlier stage. It is important to note that a laparoscopy aptitude test would not only be beneficial to abdominal surgery, but also to other specializations that depend heavily on laparoscopic skills, namely gynecology and urology.

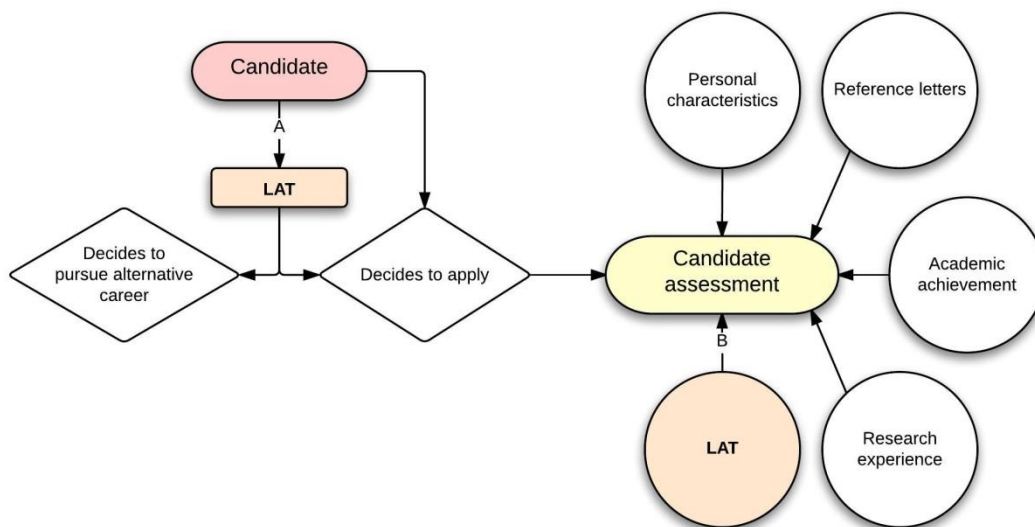


Figure 4: Contributive value of a laparoscopy aptitude test (LAT). * = Student, intern or trainee with interest in a medical career involving laparoscopic surgery. Two options for implementation. A = LAT outcome helps in the decision to pursue or not to pursue a career involving laparoscopic surgery, B = LAT outcome is used in conjunction with other factors to assess the potential of candidates for a medical specialization involving laparoscopic surgery.

Simulator-based assessment of aptitude

If the logistical and organizational burdens are perceived as acceptable and the decision is made to incorporate aptitude testing in the assessment of candidates for laparoscopic surgery, the most appropriate aptitude tests will have to be chosen. All things taken together, we would say that simulator-based assessment is the most viable option because of 2 reasons: 1) simulator-based assessment of aptitude has a relatively high correlation with future laparoscopic skills, accounting for approximately 37% - 48% ($r^2 = 0.37 - 0.48$) of the variability in performance between individuals and 2) Simulators are widely available in surgical departments involved in surgical training and can

therefore facilitate laparoscopy aptitude testing with a minimum of additional financial investment and organizational effort.

Although simulator-based assessment shows a high potency in predictive value and practical applicability, there are some important downsides that have to be mentioned. First, according to the quality assessment, all studies that included simulator-based assessment were at risk of bias. Although this does not mean the included studies were indeed all biased, the result as presented should be interpreted with caution because the quality of the studies is tangleable. Second, as the number of published cut-off scores to classify candidates according to their potential is limited and based on low sample sizes, a norm-referenced scoring system, based on the ranking of candidates, would currently be more practical than a criterion-referenced scoring system based on cut-off scores. Third, the performance level is influenced by previous training and prior experience with video games.⁵⁷ The increasing availability of home laparoscopy trainers thus creates the risk of measuring the degree of adaptation to a human-computer interface instead of assessing aptitude.

Visual-spatial, perceptual and psychomotor assessment of aptitude

The summary correlations for visual-spatial ability, perceptual ability and psychomotor ability were all statistically significant. However, calculations of the correlations in the skills lab and in the OR of the separate only account for 7% - 25% ($r^2 = 0.07 - 0.25$) of the variance in laparoscopic skills. Consequently, to obtain the laparoscopy aptitude test with the highest predictive value, visual-spatial, perceptual and psychomotor ability can best be used in combination within a laparoscopy aptitude test battery. This would increase organizational burdens, but would optimize the predictive value of a laparoscopy aptitude test by evaluating multiple aspects of potential.

Notably, correlations in this order of magnitude have not always been perceived as barriers for implementation. The correlations of pilot aptitude testing with training and flying performance after training and the correlations of the North-American dental aptitude test with practical hands-on dentistry performance levels are reported to be between 0.20 and 0.40 ($r^2 = 0.04 - 0.16$).^{54,58,59} Despite controversy around the size of these correlations, many directors of training programs in aviation and dental education have determined that aptitude testing is of contributive value and therefore continue to implement these tests in selection procedures to optimize the distribution of talent in their work force and increase training efficiency.

If the choice is made to use visual-spatial, perceptual and/or psychomotor ability instead of using simulator-based assessment, one should be aware that the majority of aptitude tests in the included studies were developed in the 4-7th decade of the 20th century (see references of Appendix D and E) and some have predominantly been evaluated in the ability to identify cognitive or psychomotor deficits in patient populations.⁶⁰ The aptitude tests are therefore not optimally adjusted to the challenges imposed by the MIS environment and it may be useful to consider composing cross-functional teams to develop new aptitude tests that reflect the demands of the MIS work environment to a higher extent than the currently available tests.

Moderator analysis

In the calculation of the summary correlation of visual-spatial ability, heterogeneity Q and the amount of variance (I^2) indicated that there was significant heterogeneity within the sample of included studies. This indicates that the size of the correlation between visual-spatial ability and laparoscopic skills might be dependent on the methodology used in the included studies. A moderator analysis was conducted to identify the differences in methodology that could have caused heterogeneity among studies.

Adjusted 2x2 classification

In the moderator analysis of visual-spatial ability, a significant difference was observed between the mean correlations of the 3D dynamic and the intrinsic static subgroup. This finding seems logical as laparoscopy requires rotation, translation and manipulation of mental structures. The significant difference in correlation could also have been caused by the majority of studies using a simulator to

measure performance level, because in general, simulator tasks do not challenge the ability of an individual to recognize objects on the basis of their characteristics. It is imaginable that in vivo laparoscopy, in contrast to simulator tasks, does require the ability to distinguish or recognize relevant structures on the basis of their intrinsic characteristics. For example, obvious visual signals to identify the cystic artery, such as a pulsation, can be absent during dissection of Calot's triangle. Surgeons must then rely on more subtle visual signals to identify the cystic artery and intrinsic static visual-spatial ability might become more relevant. Such nuances can only become visible if aptitude test scores for intrinsic static visual-spatial ability are correlated with the level of performance in these kind of subtasks. Task need analysis of laparoscopic procedures, such as described by Tjiam et al., could be used to further explore the predictive validity of the different forms of visual-spatial ability, and perhaps, also different forms of perceptual and psychomotor ability.⁶¹

Learning curve

Although the majority of studies did not decompose the correlation of aptitude with laparoscopic skills into the different phases of the learning curve, the moderator analysis of those studies that did describe correlations with baseline performance, learning rate and/or learning plateau showed that visual-spatial ability test scores significantly correlate with baseline performance, but also with learning plateau, the performance level after training. Unfortunately, correlations of visual spatial ability with learning rate are scarce and the results differ between the two studies that addressed this learning phase.^{70,74} Theoretically, there is a possibility that the learning speed in low aptitude trainees is higher than that of high aptitude trainees. Nonetheless, if we consider the finding of a higher baseline and a higher learning plateau in high aptitude trainees in this meta-analysis, it is likely that reaching proficiency goals during laparoscopic skills training requires less time for high aptitude trainees, as was found in the majority of studies that addressed this topic without distinguishing baseline performance and learning rate.^{12,13,45}

Measurement method

The summary correlations of laparoscopic skills measurements conducted with video trainers, VR simulators, laparoscopic camera navigation tasks and measurements of laparoscopic skills performed on humans or animals (subgroup OR) were all significant. The lowest correlation was observed in the subgroup video trainer and the highest in the subgroup OR. The level of complexity of the laparoscopic task and the cognitive input from the work environment probably explains the trend towards higher correlations when laparoscopic skills are measured with a performance on humans or animals (Figure 6).^{37,38}

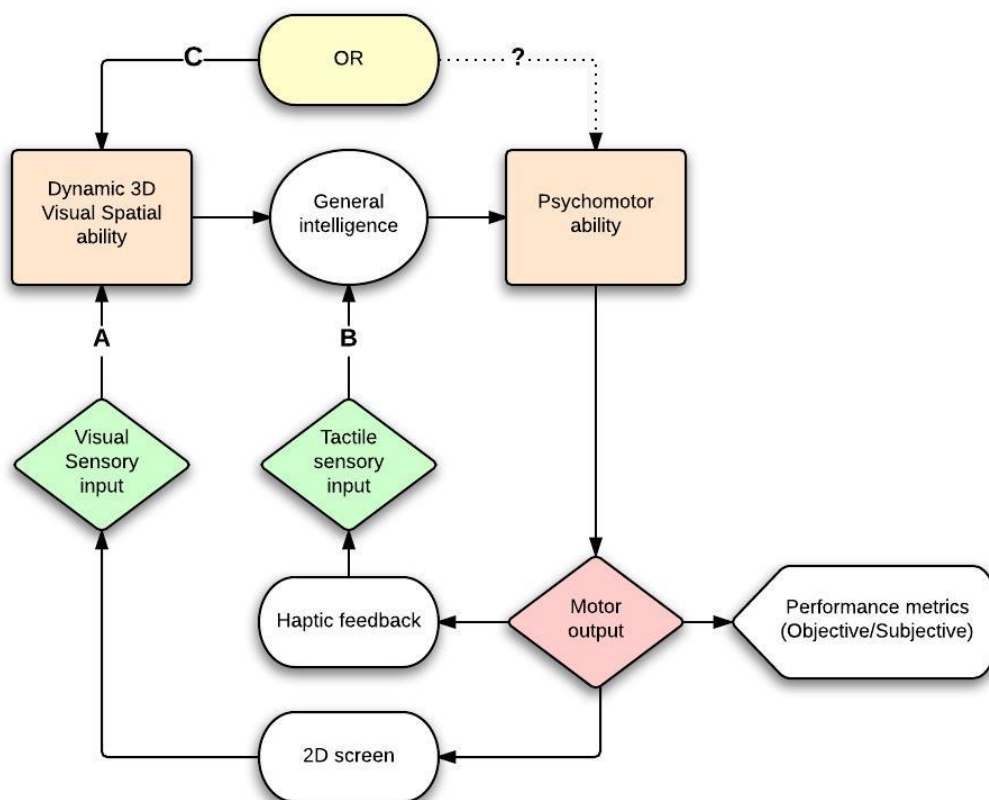


Figure 6: Flowchart of the hypothetical mechanism leading to possible differences in mean correlations between the subgroups video trainer (VT), VR simulator (VRS) and operating room (OR) in the moderator analysis of measurement method. A = visual input in subgroups VT, VRS and OR. B = additional tactile input in subgroups VT and OR, leading to the lowest mean correlation in the subgroup VT. C = high visual-spatial ability demands in the OR, whereby the influence of C > B, leading to the highest summary correlation in the subgroup OR.

Participant characteristics

Similar to the moderator learning curve, a significant correlation was found in the moderator participants characteristics in the subgroup of participants that had received laparoscopic skills training. No significant correlation was found in the subgroup non-medical-students, indicating that the inclusion of these participants can lead to results biased in the negative direction. Furthermore, a trend towards a higher correlation between visual-spatial ability and laparoscopic performance among novice trainees than among students, and a higher correlation among novice trainees than among experienced participants was observed.

There are a number of ways the recruitment of certain kinds of participants could influence the correlation in the positive or negative direction (Table 10). First, some non-medical and medical students might have a high visual-spatial ability, but not always have the affinity with manual skills needed to express this during a laparoscopic performance measurement (Figure 6). Also, the homogeneity in motivation is probably higher among novice trainees, and as a consequence, the correlation between visual-spatial ability and laparoscopic skills becomes more evident than among students.

Second, although the subgroup of experienced participants (consultant surgeons, gynecologists and trainees with training) showed a significant correlation, the correlation was lower than that of the subgroup of novice trainees. This supports the argument that laparoscopy training and experience decreases the number and frequency of unpredictable elements in laparoscopy tasks, and in doing so, decreases the association of visual-spatial ability with laparoscopic skills by having a positive effect on laparoscopic skills, but not on visual-spatial ability.

Third, the lower correlation in the subgroup of experienced participants could be caused by the rise in heterogeneity within the variable age among trained trainees and consultants. It has been shown in the field of psychology that spatial ability declines with age.⁶² In the included studies, age could function as a third variable in the moderator analysis by having a positive effect on experience, and thus laparoscopic performance measurements, and at the same time a negative effect on visual-spatial ability test scores, thereby decreasing the correlation of visual-spatial ability with laparoscopic skills.

Table 10: Research participant characteristics can potentially function as third variables in the correlation between visual-spatial ability and laparoscopic skills. The influence of third variables is more or less expelled in studies using groups of novice trainees as research participants as they are more homogeneous in psychomotor ability, motivation, age and training. As a consequence, the highest correlation is observed within these studies.

Third variable	Non-medical students	Medical students	Novice trainees	Trained participants
Affinity with psychomotor skills	*	*		
Motivation	*	*		
Age				*
Training				*
Summary correlation	Non-significant	Low	High	Low

* = negative influence on correlation between aptitude and laparoscopy within subgroup

The concept of deliberate practice

The significant correlation in the subgroup learning plateau observed in the moderator learning curve and the significant correlation in the subgroup trained participants in the moderator participants, both suggest the existence of a difference in the capability to perform laparoscopic tasks that cannot be compensated with repetitive task training. Although this indicates an innate component to task performance level after training, we have not evaluated whether there are trainees with a low aptitude score that are actually unable to learn to perform laparoscopic tasks on a proficiency level. We therefore discourage the use of these study results to harness a deterministic perspective on laparoscopic skills training. It is likely that the effects of focused 'deliberate practice', consisting of 1) training on a well-defined task, 2) detailed immediate feedback and 3) opportunities for practice tailored to individual needs, enables all candidates to eventually reach proficiency criteria in laparoscopy surgery training, and perhaps, achieve performance levels above those with high scores on aptitude assessments.⁶³ Aptitude assessment should not become a self-fulfilling prophecy, wherein those who don't perform as well on an aptitude test misattribute their inability to reach proficiency levels to a lack of talent. Motivation, perseverance and deliberate practice are greater determinants of technical performance than a test-score on aptitude tests.

Limitations

Some important limitations should be kept in mind when interpreting the results of this study. First, methodological weaknesses of this study are the risk of bias and concerns of applicability in the included studies, potential bias across studies that were used to estimate the summary correlation of perceptual ability and psychomotor ability and the possibility of an insufficient statistical power to identify significant differences in some parts of the moderator analysis of visual-spatial ability (learning curve, participant characteristics and measurement method).

Second, like in any non-experimental design, to establish that there is a causal relation between two variables, it has to be shown that the relation is not caused by the action of other variables. Two studies that addressed general intelligence as a possible confounding variable showed that the correlation with visual-spatial ability remains significant and might even increase when the correlation is corrected for general intelligence.^{37,42} Although these studies support the hypothesis of aptitude as a determinant of laparoscopic skills independent of general intelligence, further research

is necessary to identify the contribution of other confounding factors such as motivation and video game playing.

Third, some studies that addressed the advantages of binocular imaging systems have shown that the improved quality in vision is beneficial for novices and expert surgeons, inside and outside training centers.^{64,65} Thus, although 3D laparoscopy still has some considerable disadvantages, as technology develops, barriers for wide spread implementation of 3D laparoscopy might disappear and some of the findings should be reevaluated.

Fourth, medical knowledge, communication skills, decision-making skills and clinical judgment are core clinical competencies that should always be considered in conjunction with technical abilities when surgical competence is addressed. Careful selection of trainees includes a holistic perspective of competency and a thorough assessment of all technical and non-technical skills required to be a surgeon. As stated, a laparoscopy aptitude test can therefore only be considered as an additional source of information to attain a more complete picture of surgical potential.

Conclusions

In this study, the available evidence has been synthesized to provide program directors in laparoscopy related medical disciplines with the most important information for the assessment of aptitude for laparoscopic surgery among candidates. The summary correlations indicate that visual-spatial, perceptual and psychomotor ability account for part of the variance in learning and performing laparoscopic skills. Simulator-based assessment appears to have the highest predictive value by acting as a job sample wherein all aptitudes for laparoscopy are measured at once. Because of the wide availability of simulators it is also the most feasible assessment instrument. Considering the importance of technical skills in laparoscopic surgery and the current lack of methods to assess the technical potential of candidates, aptitude assessment can be of contributive value for specializations that require laparoscopic skills.

Acknowledgments

The authors would like to thank Tineke Bouwkamp-Timmer and Maarten Jalink for reviewing the manuscript. They would also like to thank Jan-Maarten Luursema, Liv Ahlborg and Jacek Sliwinski for generously providing the research data of their publications.

References

1. Schijven MP, Jakimowicz JJ, Carter FJ. How to select aspirant laparoscopic surgical trainees: establishing concurrent validity comparing Xitact LS500 index performance scores with standardized psychomotor aptitude test battery scores. *J Surg Res.* 2004;121(1):112–119.
2. Grantcharov TP, Funch-Jensen P. Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am J Surg.* 2009;197(4):447–449.
3. Bosker R, Groen H, Hoff C, Totte E, Ploeg R, Pierie J-PP. Early learning effect of residents for laparoscopic sigmoid resection. *J Surg Educ.* 2013;70(2):200–205.
4. Maan Z, Maan I, Darzi A, Aggarwal R. Systematic review of predictors of surgical performance. *Br J Surg.* 2012;99(12):1610–1621.
5. Paice A, Aggarwal R, Darzi A. Safety in Surgery: Is Selection the Missing Link? *World J Surg.* 2010;34(9):1993–2000.
6. Andriole D, Jeffe D, Whelan A. What predicts surgical internship performance? *Am J Surg.* 2004;188(2):161-164.
7. Bishawi M, Pryor A. Should technical aptitude evaluation become part of resident selection for surgical residency? *Surg Endosc.* 2014;28(10):2761-2762
8. Gallagher A, Leonard G, Traynor O. Role and feasibility of psychomotor and dexterity testing in selection for surgical training. *ANZ J Surg.* 2009;79(3):108–113.
9. Müller M. Safety lessons taken from the airlines. *Br J Surg.* 2004;91(4):393–394.
10. Groenier M, Schraagen J, Miedema H, Broeders I. The role of cognitive abilities in laparoscopic simulator training. *Adv in Health Sci Educ.* 2013;19(2):203–217.
11. Luursema J-M, Verwey W, Burie R. Visuospatial ability factors and performance variables in laparoscopic simulator training. *Learn Individ Differ.* 2012;22(5):632-638.
12. McClusky DA, Ritter EM, Lederman AB, Gallagher AG, Smith CD. Correlation between perceptual, visuo-spatial, and psychomotor aptitude to duration of training required to reach performance goals on the MIST-VR surgical simulator. *Am J Surg.* 2005;71(1):13-20.
13. Keehner MM, Tendick F, Meng MV, Anwar HP, Hegarty M, Stoller ML, et al. Spatial ability, experience, and skill in laparoscopic surgery. *Am J Surg.* 2004;188(1):71–75.
14. Louridas M, Szasz P, Montbrun S de, Harris KA, Grantcharov TP. Can We Predict Technical Aptitude? A Systematic Review. *Ann Surg.* 2015 Jun 15.
15. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–36.
16. Borenstein M, Hedges LV, Higgins J, Rothstein HR. Introduction to meta-analysis. 2011
17. Huedo-Medina T, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods.* 2006;11(2):193–206.
18. Uttal D, Meadow N, Tipton E, Hand L, Alden A, Warren C, et al. The malleability of spatial skills: a meta-analysis of training studies. *Psychol bull.* 2013;139(2):352–402.
19. Study NE. An overview of tests of cognitive spatial ability. In: 66th Engineering Design Graphic Division Mid-Year Conference Proceedings; 2012 January 22-24; Galveston, Texas, USA. Galveston: Texas; 2012. p.92-97.
20. Kozhevnikov M, Kosslyn S, Shephard J. Spatial versus object visualizers: a new characterization of visual cognitive style. *Memory & cognition.* 2005;33(4):710–26.
21. Blazhenkova O, Kozhevnikov M. Visual-object ability: A new dimension of non-verbal intelligence. 2010.
22. Feldman L, Cao J, Andalib A, Fraser S, Fried G. A method to characterize the learning curve for performance of a fundamental laparoscopic simulator task: Defining “learning plateau” and “learning rate.” *Surgery.* 2009;146(2):381386.

23. Choi Y, Kim Z, Hur K. Learning curve for laparoscopic totally extraperitoneal repair of inguinal hernia. *Can J Surg*. 2012;55(1):3336.
24. Ritter FE, Baxter GD, Kim JW, Srinivasmurthy, S. Learning and retention. In: Lee JD & Kirlik A, editors. *The Oxford Handbook of Cognitive Engineering* (pp. 125-142). New York, NY: Oxford; 2013. P. 125-142.
25. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315(7109):629-634.
26. Sterne J, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J Clin Epidemiol*. 2000;53(11):1119-1129.
27. Buckley C, Kavanagh D, Gallagher T, Conroy R, Traynor O, Neary P. Does aptitude influence the rate at which proficiency is achieved for laparoscopic appendectomy? *J Am Coll Surg*. 2013;217(6):1020-1027.
28. Buckley C, Kavanagh D, Nugent E, Ryan D, Traynor O, Neary P. The impact of aptitude on the learning curve for laparoscopic suturing. *Am J Surg*. 2013;207(2):263-270.
29. Bartenbach H. Leercurves op de LapSim in relatie tot cognitieve vaardigheden [Bsc thesis]. Tilburg, Noord-Brabant: Tilburg University; 2014; [cited 2015 Jan 7]. Available from: <http://essay.utwente.nl/65616/>.
30. Rosenthal R, Gantert WA, Scheidegger D, Oertli D. Can skills assessment on a virtual reality trainer predict a surgical trainee's talent in laparoscopic surgery? *Surg Endosc*. 2006;20(8):1286-1290.
31. Bruwaene VS, Win DG, Schijven M. Effect of a Short Preclinical Laparoscopy Course for Interns in Surgery: A Randomized Controlled Trial. *J Surg Educ*. 2014;71(2):187-192.
32. Cadeddu JA, Kondraske GV. Human performance testing and simulators. *J Endourol*. 2007;21(3):300-304.
33. Utesch T. Effects of cognitive aptitude on the initial performance on a laparoscopic simulator. [Bsc thesis]. Tilburg, Noord-Brabant: Tilburg University; 2014; [cited 2015 Jan 7]. Available from: <http://essay.utwente.nl/65809/>.
34. Hilgerink J. De rol van cognitieve vaardigheden op een laparoscopische simulator. [Bsc thesis]. Tilburg, Noord-Brabant: Tilburg University; 2014; [cited 2015 Jan 7]. Available from: <http://essay.utwente.nl/64791/>.
35. Stefanidis D, Korndorffer J, Black F, Dunne J, Sierra R, Touchard C, et al. Psychomotor testing predicts rate of skill acquisition for proficiency-based laparoscopic skills training. *Surgery*. 2006;140(2):252-262.
36. McCluney A, Vassiliou M, Kaneva P, Cao J, Stanbridge D, Feldman L, et al. FLS simulator performance predicts intraoperative laparoscopic skill. *Surg Endosc*. 2007;21(11):1991-1995.
37. Keehner M, Lippa Y, Montello D, Tendick F, Hegarty M. Learning a spatial skill for surgery: how the contributions of abilities change with practice. *Appl Cognit Psychol*. 2006;20(4):487-503.
38. Ahlborg L, Hedman L, Rasmussen C, Felländer-Tsai L, Enochsson L. Non-technical factors influence laparoscopic simulator performance among OBGYN residents. *Gynecol Surg*. 2012;9(4):415-420.
39. Ahlberg G, Heikkinen T, Iselius L, Leijonmarck C-E, Rutqvist J, Arvidsson D. Does training in a virtual reality simulator improve surgical performance? *Surg Endosc*. 2001;16(1):126-129.
40. Hogle N, Widmann W, Ude A, Hardy M, Fowler D. Does training novices to criteria and does rapid acquisition of skills on laparoscopic simulators have predictive validity or are we just playing video games? *J Surg Educ*. 2008;65(6):431-435.
41. Eyal R, Tendick F. Spatial ability and learning the use of an angled laparoscope in a virtual environment. *Stud Health Technol Inform*. 2001;81:146-152.
42. Hedman L, Ström P, Andersson P, Kjellin A, Wredmark T, Felländer-Tsai L. High-level visual-spatial ability for novices correlates with performance in a visual-spatial complex surgical simulator task. *Surg Endosc*. 2006;20(8):1275-1280.

43. Jungmann F, Gockel I, Hecht H, Kuhr K, Räsänen J, Sihvo E, et al. Impact of perceptual ability and mental imagery training on simulated laparoscopic knot-tying in surgical novices using a Nissen fundoplication model. *Scand J Surg*. 2011;100(2):78–85.
44. Ahlborg L, Hedman L, Murkes D, Westman B, Kjellin A, Felländer-Tsai L, et al. Visuospatial ability correlates with performance in simulated gynecological laparoscopy. *Eur J Obstet Gynecol Reprod Biol*. 2011;157(1):7377.
45. Nugent E. The evaluation of fundamental ability in acquiring minimally invasive surgical skill sets [MD thesis]. Dublin: Royal College of Surgeons in Ireland; 2012 [cited 2015 Jan 6]. Available from: <http://epubs.rcsi.ie/mdtheses/32/>.
46. Macmillan AI, Cuschieri A. Assessment of innate ability and skills for endoscopic manipulations by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. *Am J Surg*. 1999;177(3):274–277.
47. Ahlborg L, Hedman L, Nisell H, Felländer-Tsai L, Enochsson L. Simulator training and non-technical factors improve laparoscopic performance among OBGYN trainees. *Acta Obstet Gynecol Scand*. 2013;92(10):1194–1201.
48. Gallagher A, Cowie R, Crothers I, Jordan-Black J, Satava R. PicSOR: An objective test of perceptual skill that predicts laparoscopic technical skill in three initial studies of laparoscopic performance. *Surg Endosc*. 2003;17(9):1468–1471.
49. Kundhal PS, Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surg Endosc*. 2009;23(3):645–649.
50. Cope DH, Fenton-Lee D. Assessment of laparoscopic psychomotor skills in interns using the MIST Virtual Reality Simulator: a prerequisite for those considering surgical training? *ANZ J Surg*. 2008;78(4):291–296.
51. Langlois J, Wells G, Lecourtois M, Bergeron G, Yetisir E, Martin M. Spatial abilities of medical graduates and choice of residency programs. *Anat Sci Educ*. 2015;8(2):111–119.
52. Panait L, Larios J, Brenes R, Fancher T, Ajemian M, Dudrick S, et al. Surgical skills assessment of applicants to general surgery residency. *J Surg Res*. 2011;170(2):189–194.
53. Makdisi G, Takeuchi T, Rodriguez J, Rucinski J, Wise L. How we select our residents—a survey of selection criteria in general surgery residents. *J Surg Educ*. 2011;68(1):67–72.
54. Griffin GR, Koonce JM. Review of psychomotor skills in pilot selection research of the U.S. military services. *Int J Aviat Psychol*. 1996;6(2):125–147.
55. Hsi S, Marcia CL, John EB. The role of spatial reasoning in engineering and the design of spatial instruction. *J Eng Educ*. 1997;86(2):151–158.
56. Ackerman PL, & Cianciolo AT. Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *J Exp Psychol Appl*. 2000;6(4):259.
57. Jalink MB, Goris J, Heineman E, Pierie J-PENP, Cate Hoedemaker HO ten. The effects of video games on laparoscopic simulator skills. *Am J Surg*. 2014;208(1):151–156.
58. Ranney RR, Wilson MB, Bennett RB. Evaluation of applicants to predoctoral dental education programs: review of the literature. *J Dent Educ*. 2005;69(10):1095–1106.
59. Hegarty M, Keehner M, Khooshabeh P, Montello D. How spatial abilities enhance, and are enhanced by, dental education. *Learn Individ Differ*. 2009;19(1):6170.
60. Causby R, Reed L, McDonnell M, Hillier S. Use of objective psychomotor tests in health professionals. *Percept Mot Skills*. 2014;118(3):765–804.
61. Tjiam IM, Schout BM, Hendriks AJ, Scherpbier AJ, Witjes JA, van Merriënboer JJ. Designing simulator-based training: an approach integrating cognitive task analysis and four-component instructional design. *Med Teach*. 2012;34(10):698–707.
62. Hertzog C, Rypma B. Age differences in components of Mental-Rotation Task Performance. *Bull Psychon Soc*. 1991;29:209–212.
63. Wanzel KR, Hamstra SJ, Anastakis DJ, Matsumoto ED, Cusimano MD. Effect of visual-spatial ability on learning of spatially-complex surgical skills. *Lancet*. 2002;359(9302):230–231.

64. Lusch A, Bucur P, Menhadji A, Okhunov Z, Liss M, Perez-Lanzac A, et al. Evaluation of the Impact of Three-Dimensional Vision on Laparoscopic Performance. *J Endourol.* 2014;28(2):261-266.
65. Sahu D, Mathew M, Reddy P. 3D Laparoscopy - Help or Hype: Initial Experience of A Tertiary Health Centre. *J Clin Diagn Res.* 2014; 8:NC01-3.
66. Risucci, Geiss, Gellman, Pinard, Rosser JC. Experience and visual perception in resident acquisition of laparoscopic skills. *Current surgery.* 2000;57(4):368–372.
67. Risucci D, Geiss A, Gellman L, Pinard B, Rosser J. Surgeon-specific factors in the acquisition of laparoscopic surgical skills. *Am J Surg.* 2001;181(4):289-293.
68. Haluck RS, Gallagher AG, Satava RM. Reliability and validity of Endotower, a virtual reality trainer for angled endoscope navigation. *Stud Health Technol Inform.* 2002;85:179-184.
69. Birbas KN, Tzafestas CS, Kaklamanos IG, Vezakis AA, Polymeneas G, Bonatsos G. Spatial ability can predict laparoscopy skill performance of novice surgeons. 16th International Congress of the European Association for Endoscopic Surgery (EAES), Stockholm, Sweden, 11–14 June 2008.
70. Andalib A, Feldman LS, Cao J, McCluney AL, Fried GM. Can Innate visuospatial abilities predict the learning curve for acquisition of technical skills in laparoscopy? Scientific Session of the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), Dallas, Texas, USA, 26–29 April 2006.
71. Hassan I, Gerdes B, Koller M, Dick B, Hellwig D, Rothmund M, et al. Spatial perception predicts laparoscopic skills on virtual reality laparoscopy simulator. *Childs Nerv Syst.* 2007;23(6):685–689.
72. Enochsson L, Ahlborg L, Murkes D, Westman B, Hedman L, Kjellin A, Tsai-Felländer L. Visuospatial ability affects the performance of gynaecological simulation in the LapSimGyn® VR simulator. 16th International Congress of the European Association for Endoscopic Surgery (EAES) Stockholm, Sweden, 11–14 June 2008.
73. Sliwinski J. Visuo-spatial ability and damage in laparoscopic simulator training. [Bsc thesis]. Tilburg, Noord-Brabant: Tilburg University; 2010; [cited 2015 Jan 7]. Available from: <http://essay.utwente.nl/60117/>.
74. Kolozsvari N, Andalib A, Kaneva P, Cao J, Vassiliou M, Fried G, et al. Sex is not everything: the role of gender in early performance of a fundamental laparoscopic skill. *Surg Endosc.* 2010;25(4):1037–1042.
75. Schlickum M, Hedman L, Enochsson L, Henningsohn L, Kjellin A, Felländer-Tsai L. Surgical simulation tasks challenge visual working memory and visual-spatial ability differently. *World J Surg.* 2011;35:710-715.
76. Nugent E, Hseino H, Boyle E, Mehigan B, Ryan K, Traynor O, et al. Assessment of the role of aptitude in the acquisition of advanced laparoscopic surgical skill sets. *Int J Colorectal Dis.* 2012;27(9):1207–1214.
77. Chaudhry A, Sutton C, Wood J, Stone R, McCloy R. Learning rate for laparoscopic surgical skills on MIST VR, a virtual reality simulator: quality of human-computer interface. *Ann R Coll Surg Engl.* 1999;81(4):281–286.

PART II: TRAINING



Chapter 3

**THE PARETO-ANALYSIS FOR ESTABLISHING
CONTENT CRITERIA IN SURGICAL TRAINING**

Kelvin H. Kramp, Marc J. van Det, Nic J.G.M. Veeger, Jean-Pierre E.N. Pierie

Journal of Surgical Education 2016;73(5):892-901

Abstract

Introduction: Current surgical training is still highly dependent on expensive OR experience. While there have been many attempts to transfer more training to the skills lab, little research is focused on which technical behaviours can lead to the highest profit when they are trained outside the OR. The Pareto-principle states that in any population that contributes to a common effect a few account for the bulk of the effect. This principle has been widely used in business management to increase company profits. This study uses the Pareto-principle for establishing content criteria for more efficient surgical training.

Method: A retrospective study was conducted to assess the verbal guidance provided by 9 supervising surgeons to 12 trainees performing 64 laparoscopic cholecystectomies in the OR. The verbal corrections were documented, tallied and clustered according to the aimed change in novice behaviour. The corrections were rank-ordered and a cumulative distribution curve was used to calculate which corrections accounted for 80% of the total number of verbal corrections.

Results: In total, 253 different verbal corrections were uttered 1587 times and were categorized into 40 different clusters of aimed changes in novice behaviours. The 35 highest ranking verbal corrections (14%) and the 11 highest ranking clusters (28%) accounted for 80% of the total number of verbal corrections.

Conclusions: Following the Pareto-principle, we were able to identify the aspects of trainee behaviour that account for the majority of corrections given by supervisors during a laparoscopic cholecystectomy on humans. This strategy can be used for the development of new training programs to prepare the trainee in advance for the challenges encountered in the clinical setting in an OR.

Introduction

In 1887, the Italian economist Vilfredo Pareto observed an exponential relation between the amount of wealth an inhabitant owned and the rank-order of the inhabitant.¹ He discovered that 80% of property is owned by merely 20% of the inhabitants, a pattern which later was popularized in the 1950s by management consultant Joseph M. Juran as the Pareto-principle or '80-20 rule'.² The Pareto-principle is best known for its use in increasing business returns by identifying the vital-few causes responsible for the bulk of income within a company and consequently increasing its efficiency by focusing investments on these company facets.²⁻⁵ The Pareto-principle has also been observed in many other fields.⁶⁻⁹

In the surgical profession, the operating room is the ultimate teaching venue for learning surgical skills. However, learning how to operate costs significant amounts of money and time. Bridges and Diamond compared the operative times of cases performed by faculty with those performed by residents and calculated that the increased operative times during surgical training cost \$47,970 per year per resident.¹⁰ Furthermore, it seems that the exposure of residents to surgical procedures is decreasing because of the implementation of work hours restrictions.¹¹ These findings underline the need for higher training efficiency in the operating room (OR).

Previous studies that have described content criteria for surgical training based their findings mainly on cognitive task analysis, human reliability analysis or expert opinion.¹²⁻¹⁴ A cognitive task analysis consists of the identification of the different cognitive and procedural steps that have to be undertaken to complete a procedure.¹² Information about these steps is obtained through an interview of experts and can be used as a 'blueprint' for the development of training tasks for a procedure. Human reliability analysis has been used in high-risk technological advanced industries, such as aviation and nuclear power plant development, but has recently also been used in laparoscopic surgery as a means for developing surgical training content.¹⁵⁻¹⁷ Human reliability analysis consists of identifying what can go wrong, estimating the probability and consequences of the errors and consequently developing (training) methods to minimize the risk and consequences of these errors. While cognitive task analysis, human reliability analysis and expert opinion all provide valuable information for surgical training curriculum development, they do not provide us with a description of the aspects of surgical expertise that requires the most time and energy during training in the OR. Meanwhile, the Pareto-analysis might provide a valuable tool in the reduction of training duration in the OR by identifying those aspects of surgical skills that require the most resources to instill in trainees. This study attempts to answer the following research questions:

- 1) Does the Pareto-principle exist in the surgical training of a basic surgical procedure?
- 2) What are the content criteria for more efficient surgical training stated by means of the Pareto-principle?
- 3) How can surgical training in the dry lab and in the OR be adapted to these criteria?

Method

This study was a retrospective analysis of operative videos of laparoscopic cholecystectomies recorded for other study purposes. All the videos were recorded in Leeuwarden Medical Centre, a regional high volume teaching hospital performing >200 laparoscopic cholecystectomies per year.

Data collection

The laparoscopic cholecystectomy, a frequently performed laparoscopic training procedure, was used for the Pareto-analysis. The audio-visual recordings of laparoscopic cholecystectomies performed during two prospective studies conducted in our institution were retrospectively reviewed. The first study was conducted by van Det in 2008, the second by Kramp in 2014.^{18,19} The trainees in these videos had performed 0-20 procedures as first surgeon.

Surgical training

In both study periods, each trainee was a resident in surgery and had completed a simulator course in basic laparoscopic skills training on the SIMENDO laparoscopy trainer (Simendo, Rotterdam, the Netherlands) before commencing supervised laparoscopic surgery on patients. Knowledge of the relevant anatomy and procedural steps necessary to complete the procedure was acquired by trainees through the usual sources available online and within our institution (anatomy books, online information, example videos, etc.).

During supervised surgical training in the OR, supervising surgeons aim to find a balance between creating the optimal learning experience and guarding the patient safety during the operation. They therefore guide trainees through the procedure by giving verbal guidance and taking over when necessary while they act as assistant surgeon. The verbal guidance was divided into two different categories, verbal instructions and verbal corrections. Verbal instructions were defined as the verbal guidance provided to initiate a certain surgical behaviour (e.g. “make an incision from point a to point b”). Verbal corrections are given to reduce potentially unsafe surgical behavioural patterns or to optimize the degree of skilfulness while surgical behaviour is being exhibited by a trainee (e.g. “stay closer to the gallbladder”). Medical declarative knowledge is usually evaluated by the supervising surgeon through a quizzing behaviour, by Sutkin et al. described as ‘Socratic-like questioning to assess the surgical trainee’s knowledge’.^{20,21} While the aim of these questions is primarily to stimulate thinking about a particular aspect of the procedure, the corrections of wrong answers on these questions were also classified as verbal corrections.

Furthermore, if a supervising surgeon perceives an operative step as a particularly difficult dissection (e.g. as a consequence of variation in anatomy), or perceives the trainee as incompetent to deal with a certain aspect of the operation, he or she might temporarily take over one or both instruments to guard the flow and safety of the procedure. The exact content of the verbal guidance and reasons for a takeover are based on the supervising surgeon’s judgment of the observed situational characteristics of the operation (e.g. time pressure, anatomic variation, inflammation, etc.) and surgical behaviour of the trainee.

Evaluation of the Pareto-principle

To evaluate whether the Pareto-principle exists in the verbal corrections during surgical training, the different verbal corrections had to be counted. The data collection method used for counting verbal corrections was based on the sampling methods for observational studies of animal behaviour described by Altmann (Figure 1).²² The audio of the videos was used to document the content of the verbal corrections given by the supervising surgeon during the operation. The verbal corrections were documented in computerized sheets and were time coded. If the same novice behaviour was observed on separate occasions multiple times during a procedure, the number of repetitions of the verbal corrections to correct the behaviour was counted. If multiple verbal corrections were given to

clarify the primary verbal correction, they were counted as one verbal correction as shown in figure 2.

Verbal instructions were not counted because they are predominantly used to initiate a behaviour in the trainee (e.g. clip the artery), and therefore provide little information about what is challenging about a specific behaviour (e.g. optimal exposure of the artery by exercising traction on the gallbladder). Takeovers were not counted because the exact reason for a takeover is often not made clear by the supervising surgeon. Verbal corrections to adjust the viewing perspective of the camera when the trainee was acting as the assisting surgeon during a takeover were also not counted.

The verbal corrections were clustered according to the corrected behaviour. For instance, while using the dissection hook (behaviour), a trainee can be corrected to 'look for the silver sign', 'not to burn while applying traction with the hook', 'not to apply diathermia too close to the instrument tip of the opposite hand', etc. To identify handling of the dissection hook as the behaviour that was challenging during training these instances, the counts of these corrections were summed in the cluster 'use of the dissection hook'.

Finally, the different verbal corrections and the clusters of verbal corrections were rank-ordered on the basis of the total frequency of the verbal corrections they contained. The Pareto-principle was evaluated for the individual verbal corrections and for the clusters of verbal corrections by: 1) Plotting the number of verbal corrections as a function of the rank-order of the individual verbal corrections, 2) Plotting the number of verbal corrections within a cluster as a function of the rank-order of the clusters of verbal corrections and 3) Evaluating whether the curves showed a power law distribution with a cumulative distribution curve similar to those observed in other data.⁶

Establishing content-criteria for surgical training by means of the Pareto-principle

Whole procedure

A cut-off value of 80% was used in the cumulative distribution curve of the clustered verbal corrections to identify training content that could be used for increasing training efficiency.

Operative steps

To estimate the highest ranking corrected novice behaviours per procedural step, the start and end time of the different key steps of the recorded procedures was determined according to a previously published study.²³ The procedural steps include: (1) open introduction of the first trocar and accessory trocar placement, (2) opening of the peritoneal envelope, (3) creating critical view of safety, (4) clipping and division of cystic duct and artery, (5) retrograde cholecystectomy, and (6) gallbladder removal and closure.

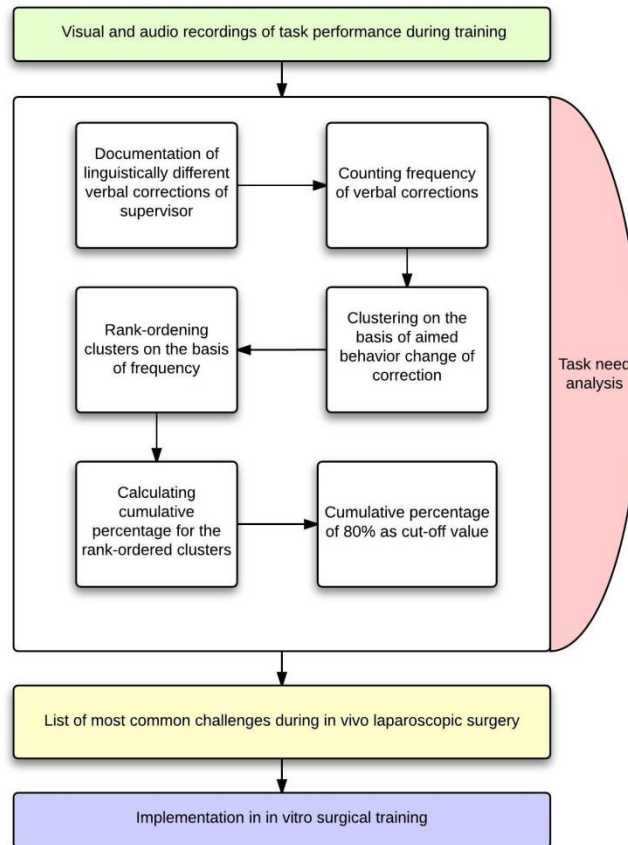


Figure 1: Flowchart of study methodology.

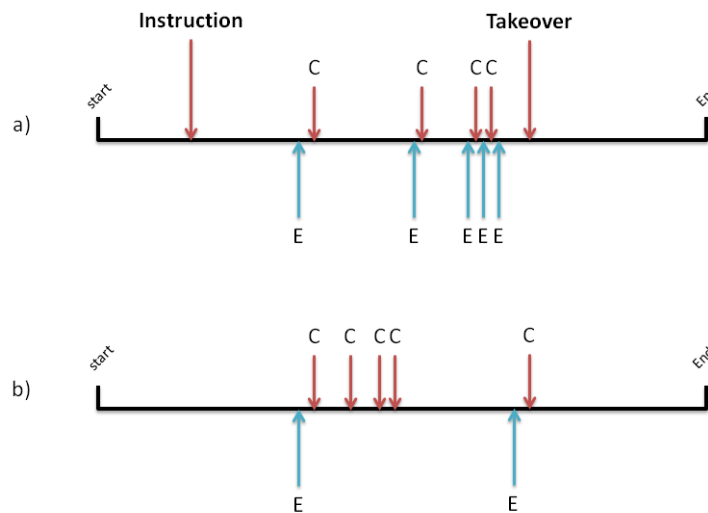


Figure 2: Counting of verbal corrections: red = communication from supervising surgeon to trainee, blue = erroneous or unskilled behavior of trainee, C = Verbal Correction, E = Expressed erroneous or unskilled surgical behavior. **a)** Instruction is given by supervisor (e.g. dissect gallbladder from liver bed by opening the peritoneum at the right of the gallbladder). The trainee portrays a behavior that requires an increasing frequency of supervisory corrections. After 5 times this behavior has been observed, the supervisor takes over to proceed the operation. Instruction and takeovers are not counted as supervisory correction, thus $N_{vc} = 4$ in this case. **b)** The first verbal correction is not sufficient to correct the trainee because of confusing communication. A number of verbal corrections with the intention to clarify the primary correction are given. In this case $N_{vc} = 2$.

Observed perceived importance

If one surgeon of the included surgeons considers one kind of technical surgical behaviour more important than other surgeons, this could potentially bias the results. To evaluate whether this was the case, the 'observed perceived importance' was calculated for the supervising surgeons with:

$$\text{Observed perceived importance} = [N_{vc-ss}/(N_{proc-ss} \times N_{total-ss})] / [\sum (N_{vc-ssi}/(N_{proc-ssi} \times N_{total-ssi}))] \quad (1)$$

Where N_{vc-ss} = the total number of verbal corrections in one cluster given by a supervising surgeon, $N_{proc-ss}$ = total number of procedures wherein the surgeon acted as the supervisor (correction for number of supervised procedures) and $N_{total-ss}$ = total number of verbal corrections of a supervising surgeon over all clusters (correction for talkativeness). The number of verbal corrections per cluster normalized for number of supervised procedures and talkativeness was divided by the sum of the normalized numbers of verbal correction of all surgeons to obtain a percentage. Because this method is specifically aimed to screen for surgeon specific bias in the results of a Pareto-analysis, no threshold values were available in the literature to guide interpretation. Consequently we determined threshold values as follows: 1) To minimize sampling error, only surgeons with >10 supervised procedures were included and 2) An absolute difference of 30% between the maximum and minimum observed perceived importance was used as a cut-off value for identifying differences in supervisorial behaviour between surgeons.

Results

Data characteristics

A total of 64 procedures performed by 12 trainees and supervised by 9 surgeons were analyzed. The median number of videos wherein a surgeon acted as a supervisor was 4 [range 1-19.5] (in one video, the supervising surgeon had to leave in the middle of a procedure and another one took over supervision). The median number of procedures performed by the trainees in the videos was 4 [range 1-8].

Evaluation of the Pareto-principle

The videos contained 1587 verbal corrections in total. A rank-ordered distribution of the counts of the verbal corrections is shown in figure 3. Eighty percent of the total number of verbal corrections was caused by 35 of the 253 different corrections (14%).

The verbal corrections were categorized in 40 different clusters of technical behaviour (Appendix F). Fourty-six times a verbal corrections was categorized in more than one cluster. Eighty percent of the total number of verbal corrections within the clusters was caused by the 11 highest ranking clusters (28%) (Figure 4).

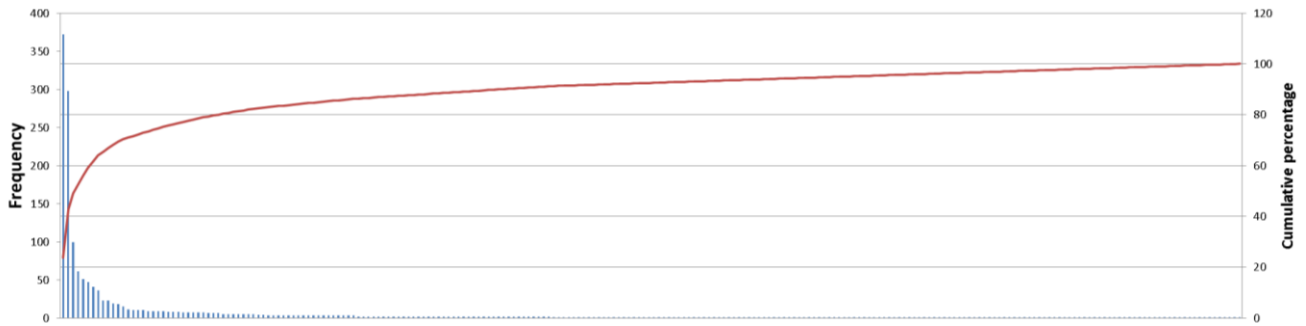


Figure 3: Rank-ordered counts (blue) and cumulative distribution curve (red) of the 253 different verbal corrections.

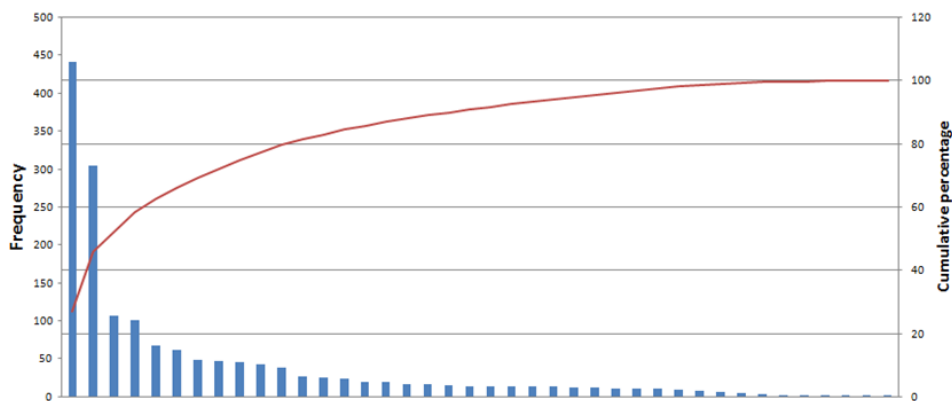


Figure 4: Rank-ordered counts (blue) and cumulative distribution curve (red) of the 40 clusters of novice behavior.

Establishing content-criteria for surgical training by means of the Pareto-principle

Whole procedure

A list of the 11 highest ranking clusters is shown in table 1. In the whole procedure, the cluster 'tensioning the gallbladder with the appropriate direction and strength' was the cluster with the highest number of corrections, accounting for 27.0% of clustered corrections. The cluster 'identifying the correct surgical plane' accounted for 18.6% of clustered corrections.

Operative steps

For the classification of verbal corrections into steps, steps 2 and 3 were merged because the operating team frequently shifted between the sub steps of these two procedural steps (e.g. opening the peritoneum and performing dissection of Calot's triangle at the left side of the gallbladder before proceeding to the peritoneum at the right side). The percentage of verbal corrections in step 1 was 9.4 %, step 2-3 was 42.2%, step 4 was 11.4%, step 5 was 32.5% and step 6 was 4.5%. The clusters of verbal corrections with the highest frequency was choosing the position and direction of the trocars (66.9%) in step 1, applying left-hand traction on the gallbladder with the appropriate strength and direction in step 2-3, step 4 and step 5 (resp. 25.6%, 43.2% and 34.9%), and using the endobag in step 6 (38.0%). The second most frequent cluster of verbal corrections were making an incision in step 1 (13.2%), determining the optimal direction of dissection in step 2-3 and 5 (resp. 21.8% and 28.9%), the use of the clipping instrument in step 4 (25.7%) and the use of the crocodile clamp in step 6 (23.9%).

Observed perceived importance

To evaluate whether one of the supervising surgeons considered one of the clusters as more important or less important than other surgeons, observed perceived importance was calculated for the verbal corrections. In the clusters 'Staying close to the gallbladder,' 'Staying superficial during dissection,' 'Use of the clipping instrument,' 'Avoiding liver damage,' and 'Positioning of the clip' the differences between surgeons exceeded the threshold of 30% (Table 1).

Table 1: Overview of clusters of verbal corrections contributing 80% of the total number of corrections given by clinical supervisors and the key steps in which the behaviors were corrected. 1) Open introduction of the first trocar and accessory trocar placement, (2) opening of the peritoneal envelope, (3) creating the CVS, (4) clipping and division of cystic duct and artery, (5) retrograde cholecystectomy, and (6) gallbladder removal and closure.

Order nr.	Verbal correction	N	f _{procedure}	Step 1	Step 2-3	Step 4	Step 5	Step 6	Min.	Max.	Diff
1	Tensioning the gallbladder with the appropriate direction and strength	441	6.89	-	+	+	+	-	24,6	43,6	19,0
2	Identifying the correct surgical plane	304	4.75	-	+	-	+	-	25,0	48,9	23,8
3	Use of the dissection hook	106	1.66	-	+	-	+	-	30,1	38,1	8,0
4	Choosing position and direction of trocar placement	101	1.58	+	-	-	-	-	28,1	43,2	15,1
5	Using the clamp	67	1.05	-	+	-	+	-	19,8	43,9	24,0
6	Staying close to the gallbladder	61	0.95	-	+	-	+	-	19,3	55,3	35,9
7	Staying superficial during dissection	49	0.77	-	+	-	+	-	15,4	61,4	45,9
8	Using the clipping instrument	47	0.73	-	-	+	-	-	17,7	55,5	37,8
9	Avoiding harm to surrounding structures other than the liver	45	0.70	-	+	-	+	-	19,4	48,8	29,4
10	Avoiding liver damage	42	0.66	-	+	-	+	-	18,8	56,3	37,4
11	Positioning of the clip	39	0.61	-	-	+	-	-	18,7	49,9	31,2
1 - 40		1633*		+	+	+	+	+	-	-	-

+ = correction has been addressed in the operative step, - = correction has not been addressed in the operative step.

*= 46 times a verbal correction was categorized in more than one cluster.

Discussion

For a job training to be useful, the appropriate training content must be identified through a proper analysis of job requirements. Surgery is a psychomotor and cognitive challenging discipline. Many different sensory motor patterns and cognitive schemata have to be acquired to perform surgery independently in a safe and skilful manner. Intuitively, the available training resources would be distributed among the spectrum of necessary skills to become a competent surgeon. However, if there is a misbalance during training in favour of certain surgical behaviours in the OR, it would be more profitable to prioritize investments in the training of those specific surgical skills. The Pareto-principle is a well-established theory in business management and states that the bulk of a common effect is caused by just a few of the causes. It is used to increase business returns by investing company resources in those aspects of business that have the highest revenue. In this study we evaluated whether the Pareto-principle is true for OR training in a basic surgical procedure, the laparoscopic cholecystectomy. Secondly, the verbal guidance expressed by supervising surgeons during training was analyzed with the Pareto-analysis in order to identify training content that could be used to increase the efficiency of training in a basic surgical procedure.

Does the Pareto-principle exist in the surgical training of a basic surgical procedure?

The separate and clustered verbal corrections plots showed a power law distribution with a cumulative distribution curve similar as those observed in other studies.⁶ Furthermore, 35 (14%) of the 253 different verbal corrections and 11 (28%) of the 40 clusters of novice behaviour accounted for 80% of the corrections given by supervisors, confirming the 80-20 rule. Based on these findings, it seems that the Pareto-principle can be demonstrated in the verbal corrections uttered by supervising surgeons during a surgical procedure.

Establishing content-criteria for surgical training by means of the Pareto-principle

The next step is to develop training methods for the job-requirements identified with the Pareto-principle. In general, these training methods could consist of all educational resources currently available to trainees such as textbook explanations, educational videos, instructional courses, dry lab training tasks, etc. We have chosen to specifically discuss training tools for the following 5 themes: 1) Tissue exposure, 2) Surgical dissection plane, 3) Instrument handling, 4) Insertion of trocars, 5) Use of the endobag.

Tissue exposure

Applying traction in the appropriate direction and with the appropriate force accounted for 27.0% of clustered corrections expressed by the supervising surgeons in our study. This is in line with an interview of experts about the most common problem areas experienced by novice trainees. These experts identified neglect of the non-dominant hand as 1 of the 5 most common difficulties.¹⁴ Although there are a number of tasks on the SIMENDO to train the non-dominant hand our results provide evidence that the content of these tasks do not suffice for adequate training of left hand coordination during a laparoscopic cholecystectomy. The full procedure simulator LapSIM, almost 20 times as expensive as the SIMENDO²⁴, includes the simulator task dissection of Calot's triangle and removal of the gallbladder from the liver bed. These are two of the operative steps in which adequate exposure of the gallbladder with the left hand is essential. Surprisingly, no explicit measures are included to assess whether the trainee adequately exposes the gallbladder through exercising traction with the right strength and in the right direction.²⁵ Horeman et al. have described a training tool to more comprehensively teach this skill. In their studies, they have demonstrated an improved tissue handling when trainees receive laparoscopic skills training with visual feedback of the size and the direction of the force they exercise through the surgical instruments.^{26,27} Therefore, a learning module wherein the right amount of force, defined by the force exerted by experts on real tissue, and direction of traction with the non-dominant hand result in the optimal exposure for

performance of a task with the dominant hand could be used to address the issue of adequate tissue exposure in surgical training.

Surgical dissection plane

Choosing the direction of dissection showed the second highest frequency of verbal corrections, accounting for 18.6% of the total number of verbal corrections. This behaviour is probably technically challenging because it consists of a complex interaction between the motor task of adequately exposing the tissue and the visual perceptual task of identifying the accurate dissection plane during exposure. Although engineers should pursue incorporating the training of this task in VR simulator tasks, a VR environment might currently not be the most suitable method for learning this behaviour due to the complexity of the tissue that needs to be simulated. There are two alternatives (other than the use of cadavers or animals) that could support training in identifying the surgical dissection plane.

A tool to transfer training in identifying the surgical dissection plane to outside the OR could be the recently validated surgical planes perception task developed by Schlachta et al.²⁸ They developed a task in identifying the accurate dissection plane in colorectal surgery by challenging subjects to draw the plane for dissection on a digital picture and calculating the distance of the line with the average line drawn by certified colorectal surgeons. A significant difference was observed between the variation in line distances among novice trainees compared with the variation among consultant surgeons for a number of the digital pictures. However, more research is currently being conducted to evaluate whether this task can actually be used to train subjects in identifying the right plane for dissection.

A second option includes technical adjustments in the OR environment to support the teaching of this topic to surgical trainees. For instance, in our institution, a trainee had once placed a marking at the middle of the screen. This allowed the supervising surgeon to point out the exact expected route of dissection for the trainee while holding the camera. A clearer visualization of dissection plane can facilitate in proceeding through the dissection a longer distance without verbal guidance than otherwise would be possible, consequently, increasing the autonomy of the trainee. Ideally, the supervising surgeon would be able to switch on a digital pointer built in the laparoscope to show the right path when the trainee loses sight on the optimal plane of dissection.

Instrument handling

The use of the dissection instruments also accounted for a high number of verbal corrections. Use of the dissection hook was ranked 3rd in the final rank-order of the clusters. Interestingly, in our study, 51 of the 106 corrections (48.1%) in this cluster were given to teach the trainee a specific pattern, namely a pull-cautery-pull pattern that consists of: 1) placing the tissue under tension by pulling, 2) activating the cautery without pulling, 3) deactivate the cautery and 4) pulling again. This pattern can be measured with measures of psychomotor skills and therefore seems a viable option for inclusion in a virtual reality or videotrainer training task. The same holds for going into the tissue parallel to the dissection plane and pulling orthogonal to the dissection plane with the hook, which accounted for 36 of the 106 (34.0%) of the verbal corrections related to the use of the dissection hook.

Trocar insertion

Choosing the correct location and direction for insertion of the trocars accounted for 66.9% of the total number of corrections in step 1. Because the use of excessive force was the most commonly cited malpractice in relation to trocar insertion²⁹, the development of the first training task dedicated to the practice of trocar insertion was focused on the number of turns needed to insert the trocars and the plunge depth during insertion.^{30,31} However, the majority of supervision is actually focused towards getting the trocar in the right location and direction instead of correcting the amount of turns or preventing too deep of a plunge. Although it might be difficult to incorporate trocar insertion in VR simulator training because of the strong dependence on haptic feedback during insertion, the variety of the abdominal wall characteristics and the preference of the surgeon, our

results suggest that future educational developments, such as textbook explanations, dedicated courses and educational videos, should preferably also include determining the (patient-specific) direction and position for insertion of trocars.

Use of the endobag

The use of the endobag did not belong to the 11 highest ranking clusters, nonetheless, corrections for the use of the endobag accounted for 38.0% of the corrections in step 6. Corrections were given during manipulation of the endobag, placing the gallbladder in the endobag and during specimen retrieval to increase the efficiency in the use of the limited intra-abdominal space. The lack of training in these skills could be, in part, related to the high expenses of the endobag (60.33 GBP/92.00 USD).³² The literature describes the use of 2 practical and inexpensive alternatives for specimen retrieval that can be used for training: 1) Tural&Schier have demonstrated the use of the innermost plastic wrapping of a Redon drain bag (0.20 GBP/0.30 USD, including Redon drain) grasped with a 2-mm needle holder and inserted through a 5-mm trocar as an alternative specimen retrieval system in children³² and 2) Yao et al. reported the extraction of 2 large gastric phytobezoars with a simple surgical glove (0.46GBP/0.70 USD) as a specimen retrieval system.³³ These alternatives could enable the addition of specimen retrieval training to the already existing training tasks on a video trainer and consequently decrease the energy and time needed to teach the use of the endobag during training in the OR.

Observed perceived importance

The variation in the observed perceived importance did not reach the threshold in the highest ranking clusters, negating the possibility that the professional judgment of one surgeon was overtly focused towards one aspect of surgical behaviour and thereby influenced their rank. As the number of verbal corrections per cluster decreases towards the lower ranking clusters, the variation in the observed perceived importance per surgeon reaches the threshold in the 6th, 7th, 8th, 9th, 10th and 11th cluster, most likely as a consequence of sampling error.

Limitations

Some limitations should be kept in mind when interpreting the results of this study. Methodological limitations include the retrospective nature of the study, a large dispersion in the number of supervised procedures per surgeon, the videos originating from studies performed in one institution coding of the audio recordings performed by one author and the subjectivity of the interpretation of interactions between persons. We also did not attempt to track the time records of the takeovers, which would have allowed a ratio of behaviour per time unit calculation instead of per procedure, one of the methods proposed by Altmann to more accurately determine behaviour during animal observation.²² This could have been a more reliable way of documenting supervisory behaviour, however the number of corrections given per time period the trainee holds the instruments in his hands is also dependent on other factors than the interaction between the professional judgment of the supervising surgeon and the observed skills of the trainee. Time pressure, patient characteristics and even the mood of the surgeon, are all factors that cannot be controlled in a retrospective study and could therefore have been factors that influenced the behaviour per time unit.

To evaluate whether surgeon specific factors have a significant influence on the study results the observed perceived percentage was calculated. However, this is a novel method that has not been validated previously. Consequently, there is no scientific evidence to support the decision to include only surgeons with a number of supervised procedure >10 and to define a difference of 30% as a cut-off point. Nonetheless, we believe that screening for surgeon specific factors by calculating a observed perceived percentage should be included in the evaluation of a Pareto-analysis as described in this study as it gives information about the generalizability of the study findings to supervising surgeons in general.

Another important methodological limitation is that this study was focused on one procedure. To confirm that this Pareto-principle holds for surgical procedures in general and to

identify learning points that can be used to increase the training efficiency of the whole scope of surgical procedures, other procedures will also have to be analysed according to the Pareto-principle. Furthermore, the data acquisition process in this study was labour intensive, raising the question on how many procedures have to be analyzed to observe an exponential pattern. Particularly in more advanced laparoscopic procedures an exponential pattern in a rank-ordered list might be hard to identify due to adjustments in behaviour and/or gains of knowledge during experience in more basic laparoscopic procedures.

Finally, it is important to note that a popular synonym for the Pareto-principle 'the vital-few and trivial many', does not hold in surgical training, as seldom corrected behaviours are not per se unimportant novice behaviours. The goal of the Pareto-analysis is to describe important aspects for increasing training efficiency, not to describe the most important aspects of the procedure itself.

Conclusion

The OR is an expensive teaching venue. Health institutions are under pressure to increase patient safety and reduce the financial costs for training in the OR. The Pareto-principle states that a few causes are responsible for the bulk of a common effect. This principle has been used within varying industries to increase business returns by increasing work process efficiency. In this study, the Pareto-principle was evaluated as a tool for the development of training content for more efficient surgical training in the OR. The verbal corrections uttered by supervising surgeons in the OR were used to explore surgical behaviours that could be the focus for better OR preparation. We found that the majority of verbal corrections were directed towards a few novice behaviours in the OR. The next step would be to validate the Pareto-principle by exploring if adequately addressing the identified behaviours in trainee preparation leads to the expected reduction in resources for health institutions.

Acknowledgments

The authors would like to thank Maarten Jalink and Ebi Cocodia for reviewing this manuscript.

References

- 1 Schumpeter JA. Vilfredo Pareto (1848-1923). *Q J Econ.* 1949;63:147–173.
- 2 Juran JM, Godfrey AB. Juran's quality handbook. 5th ed. 1999; New York: McGraw-Hill.
- 3 Ryan, T. P. Statistical methods for quality improvement. 3th ed. 2011; New Jersey: John Wiley & Sons, Inc.
- 4 Mengesha Y, Singh AP, Yimer W. Quality improvement using statistical process control tools in glass bottles manufacturing company. *Int. J. Qual. Res.* 2013;7(1):107–126.
- 5 Ahmed M, Ahmad N. An Application of Pareto Analysis and Cause-and-Effect Diagram (CED) for Minimizing Rejection of Raw Materials in Lamp Production Process. *Manag Sci Eng.* 2011;5(3):87–95.
- 6 Clauset A, Shalizi CR, Newman MEJ. Power-Law Distributions in Empirical Data. *SIAM review.* 2009;51(4):661-703.
- 7 Small M, Singer JD. Resort to arms: International and civil wars, 1816-1980. Sage Publications, Inc; 1982.
- 8 Lu ET, Hamilton RJ. Avalanches and the Distribution of Solar-Flares. *Astrophys J.* 1991;380:L89–L92.
- 9 Zipf GK. Human Behaviour and the Principle of Least Effort. Cambridge MA edn. 1949.
- 10 Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. *Am J Surg.* 1999;177:28–32.
- 11 Ahmed N, Devitt KS, Keshet I, Spicer J, Imrie K, Feldman L, et al. A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. *Ann Surg.* 2014;259(6):1041-1053.
- 12 Tjiam IM, Schout BM, Hendriks AJ, Scherpbier AJ, Witjes JA, Van Merriënboer JJ. Designing simulator-based training: An approach integrating cognitive task analysis and four-component instructional design. *Med Teach.* 2012;34(10):698-707.
- 13 Tang B, Hanna GB, Cuschieri A. Analysis of errors enacted by surgical trainees during skills training courses. *Surgery.* 2005;138(1):14-20.
- 14 Greco EF, Regehr G, Okrainec A. Identifying and Classifying Problem Areas in Laparoscopic Skills Acquisition: Can Simulators Help? *Acad Med.* 2010;85(10):S5-S8.
- 15 Joice P, Hanna GB, Cuschieri A. Errors enacted during endoscopic surgery--a human reliability analysis. *Appl Ergon.* 1998;29(6):409-414.
- 16 Cuschieri A, Tang B. Human reliability analysis (HRA) techniques and observational clinical HRA. *Minim Invasive Ther Allied Technol.* 2010;19(1):12-17.
- 17 Tang B, Hanna GB, Bax NMA, Cuschieri A. Analysis of technical surgical errors during initial experience of laparoscopic pyloromyotomy by a group of Dutch pediatric surgeons. *Surg Endosc.* 2004;18(12):1716-1720.
- 18 Van Det MJ, Meijerink WJHJ, Hoff C, Middel LJ, Koopal S, Pierie JPEN. The learning effect of intraoperative video-enhanced surgical procedure training. *Surg Endosc.* 2011;25(7):2261-2267.
- 19 Kramp KH, van Det MJ, Veeger NJGM, Pierie J-PEN. Validity reliability and support for implementation of independence-scaled procedural assessment in laparoscopic surgery. *Surg Endosc.* 2015 Sep 28. [Epub ahead of print]
- 20 Sutkin G, Littleton EB, Kanter SL. How Surgical Mentors Teach: A Classification of In Vivo Teaching Behaviors Part 1: Verbal Teaching Guidance. *J Surg Educ.* 2015;72(2):243-250.
- 21 Sutkin G, Littleton EB, Kanter SL. How Surgical Mentors Teach: A Classification of In Vivo Teaching Behaviors Part 2: Physical Teaching Guidance. *J Surg Educ.* 2015;72(2):251-257.
- 22 Altmann J. Observational study of behavior: sampling methods. *Behaviour.* 1974;49:227–267.
- 23 Bethlehem MS, Kramp KH, van Det MJ, ten Cate Hoedemaker HO, Veeger NJGM, Pierie JPEN. Development of a Standardized Training Course for Laparoscopic Procedures Using Delphi Methodology. *J Surg Educ.* 2014;71(6):810-816.

- 24 van Empel PJ, van der Veer WM, van Rijssen LB, Cuesta MA, Scheele F, Bonjer HJ, et al. Mapping the Maze of Minimally Invasive Surgery Simulators. *J Laparoendosc Adv Surg Tech A*. 2012;22(1):51-60.
- 25 Bruwaene S Van, Schijven MP, Miserez M. Assessment of Procedural Skills Using Virtual Simulation Remains a Challenge. *J Surg Educ*. 2014;71(5):654-661.
- 26 Horeman T, van Delft F, Blikkendaal MD, Dankelman J, van den Dobbelsteen JJ, Jansen F-W. Learning from visual force feedback in box trainers: tissue manipulation in laparoscopic surgery. *Surg Endosc*. 2014;28(6):1961-1970.
- 27 Rodrigues SP, Horeman T, Sam P, Dankelman J, van den Dobbelsteen JJ, Jansen F-W. Influence of visual force feedback on tissue handling in minimally invasive surgery. *Br J Surg*. 2014;101(13):1766-1773.
- 28 Schlachta C, Ali S, Ahmed H, Eagleson R. A novel method for assessing visual perception of surgical planes. *Can J Surg*. 2015;58(2):87.
- 29 Fuller J, Ashar BS, Carey-Corrado J. Trocar-associated injuries and fatalities: An analysis of 1399 reports to the FDA. *J Minim Invasive Gynecol*. 2005;12(4):302-307.
- 30 Arulesan V, Srimathveeravalli G, Kesavadas T, Nagathan P, Baier RE. Data acquisition and development of a trocar insertion simulator using synthetic tissue models. *Stud Health Technol Inform*. 2007;125:25-27.
- 31 Brümmer V, Carnahan H, Okrainec A, Dubrowski A. Trocar insertion: the neglected task of VR simulation. *Stud Health Technol Inform*. 2007;132:50-52.
- 32 Tural S, Schier F. The Use of a Plastic Bag From a Drain Package Instead of an Endobag in Children: A Safe, Effective, and Economical Alternative. *Surg Innov*. 2010;17(3):269-272.
- 33 Yao CC, Wong HH, Chen CC, Wang CC, Yang CC, Lin CS. Laparoscopic removal of large gastric phytobezoars. *Surg Laparosc Endosc Percutan Tech*. 2000;10(4):243-245.

Chapter 4

**ERGONOMIC ASSESSMENT OF THE FRENCH AND
AMERICAN POSITION FOR LAPAROSCOPIC
CHOLECYSTECTOMY IN THE MIS SUITE**

Kelvin H. Kramp, Marc J. van Det, Eric R. Totte, Christiaan Hoff, Jean-Pierre E.N. Pierie

Surgical Endoscopy 2014;28(5):1571-1578

Abstract

Aims: The cholecystectomy was one of the first surgical procedures to be performed with laparoscopy in the 1980s. Nowadays, there are generally two operation setups to perform a laparoscopic cholecystectomy: the French and the American position. In the French position the patient lies in the lithotomy position, while in the American position the patient lies supine with the left arm in abduction. In order to find an ergonomic difference between the two operation setups the movements in the vertebral column of the surgeon were analyzed in this crossover study.

Methods: The posture of the surgeon's vertebral column was recorded intra-operatively using an electromagnetic motion tracking system with three sensors attached to the head and to the trunk at the level of Th1 and S1. A three-dimensional posture analysis of the cervical and thoracolumbar spine was conducted on 4 surgeons performing a laparoscopic cholecystectomy in the French and in the American position. The body angles that were assessed consisted of: flexion/extension of the cervical and the thoracolumbar spine, axial rotation of the cervical and thoracolumbar spine, lateroflexion of the cervical and thoracolumbar spine and the orientation of the head in the sagittal plane. For each body angle, the mean, the time percentage within an ergonomic acceptable range and the relative frequencies were calculated and compared.

Results: No statistical differences were observed in the mean body angles and time percentages within an acceptable range between the French and the American position. The relative frequencies of the body angles might indicate a trend towards slight cervical flexion in the American position and slight thoracolumbar flexion in the French position.

Conclusion: In a modern dedicated minimally invasive surgery suite, there were no significant differences in body posture of the neck and trunk and orientation of the head between the French and American position.

Introduction

Since the late 1980s, cholecystectomy has been performed with a laparoscopic technique, and this currently is the gold standard. Laparoscopic surgery has several established advantages including less blood loss, decreased post-operative pain, a shorter hospital admission time, quicker reintroduction into society, and superior cosmetic results.¹⁻⁴ On the other hand, laparoscopic techniques confront the surgeon and the surgical team with ergonomic challenges. During laparoscopy, the surgeon works with a diversion of the working field and line of vision. This diversion of the visual and working axis can create awkward static postures including rotation of the spine, extension of the neck, and elevation of the upper extremities and might compromise surgical task performance.⁵⁻⁸ In recent research, approximately 87% of surgeons involved in laparoscopy reported musculoskeletal problems.⁵

Ergonomic studies suggest that a balance should be maintained between optimal comfort and safety on one hand and optimal effectiveness and efficiency on the other hand.⁹ To achieve this, the operating room has to be set up and the patient has to be positioned such that these conditions can be accommodated.⁹⁻¹⁰ For the laparoscopic cholecystectomy, two setups are widely used worldwide: the so-called French position and the American position (Figure 1). The preferred setup of surgeons is based on locoregional common practice. This study was conducted to compare body posture differences among surgeons performing a laparoscopic cholecystectomy in the French and American position.

Materials and Methods

Study design

The ergonomic qualities of the surgeon's posture in the French and American position were compared during laparoscopic cholecystectomy in a crossover design. An intraoperative motion analysis was performed during laparoscopic cholecystectomies for patients with symptomatic uncomplicated gallbladder disease.

Participating surgeons

Four surgeons (2 residents and 2 consultants) were recruited to perform the procedures in both setups (Table 1). The residents were in their 5th and 6th years of their surgical training, performing laparoscopic cholecystectomy frequently and independently.

The consultants were certified gastrointestinal surgeons with extensive experience in laparoscopic techniques. One consultant and one resident, originally trained in the Netherlands using the American position, were educated to perform laparoscopic cholecystectomy in the French position. The remaining two surgeons, originally trained in Belgium using the French position, were educated in the American position. Each of the four participants were required to perform one procedure in each position. All the surgically treated patients gave informed consent

Table 1: Education and level of experience of the participants

Surgeon	Education	Level of experience
A	American	Resident
B	American	Consultant
C	French	Resident
D	French	Consultant

Operative setup

All procedures were performed in a dedicated minimally invasive surgery (MIS) suite with permanently installed multiple flat-screen monitors attached to a ceiling-mounted suspension system. The monitor and operation table were organized to create an ergonomic workspace. The monitors were positioned according to the following guidelines^{9,10}:

1. Straight in front of the subject in the horizontal plane to avoid rotation of the vertebral column.
2. In a downward viewing direction between 10° and 30° in the sagittal plane to optimize task performance and at the same time prevent fatigue of the neck muscles.
3. At a proper viewing distance (80–120 cm), close enough to avoid loss of detail and at the same time far enough to avoid eye strain due to constant accommodation.

The table was positioned between 70% and 80% of the elbow height of the surgeon to avoid extreme excursions of the upper extremities.¹⁰

For the French position, the patient is placed in the supine position with the perineum at the edge of the table, the hips and knees flexed, and the left arm or both arms in abduction. The operating surgeon stands between the legs, the assisting surgeon standing on the right side of the patient and the scrub nurse standing on the left side (Figure 1a). The patient is turned in reversed Trendelenburg position.

For the American position, the patient also is placed in the supine position, with the left arm or both arms in abduction. The operating surgeon stands on the left side of the patient, with the scrub nurse on the left side of the operating surgeon and the assisting surgeon on the right side (Figure 1b). The patient is turned in reversed Trendelenburg position and slightly to the left. For both positions, a four-port technique is used. The optical (primary) port is located at the umbilicus. The two operating (secondary) ports are inserted at locations that enable a manipulation angle of 60

degrees between the tips of the instruments to imitate the natural relationship between the hands as far as possible. The axis of the camera is placed between the axes of the working instruments.¹¹ As a consequence of the surgeon's change in the location between the two operation positions the instrument port location is different between the two operation setups (Figure 1).

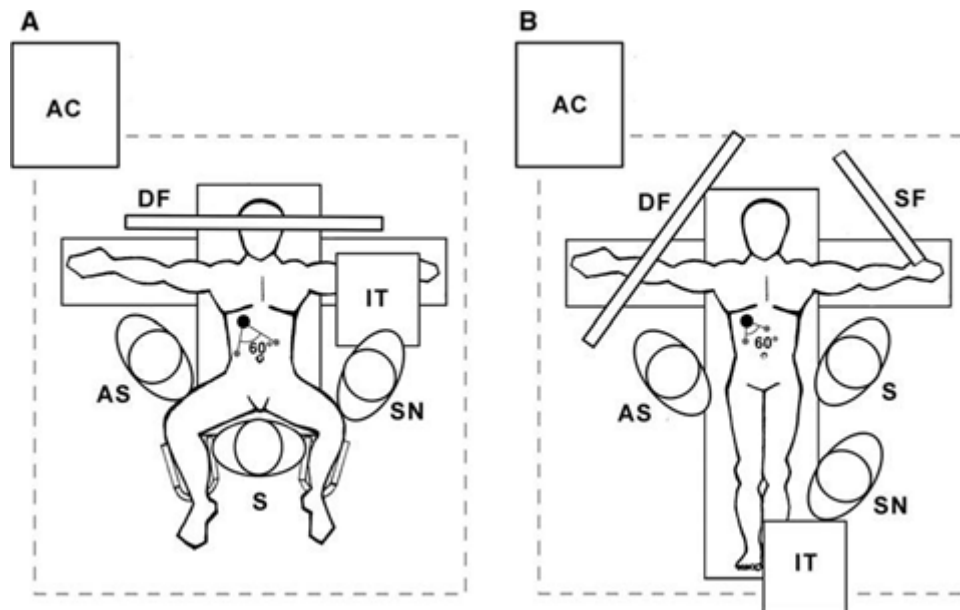


Figure 1: Room setup in dedicated minimally invasive surgery (MIS) suites with suspended monitors. a) The French position. b) The American position. AC: anesthesia console, DF: double flat screen, SF: single flat screen, S: operating surgeon, AS: assisting surgeon, SN: scrub nurse, IT: instrument table. Black dot: Location of the gallbladder. Gray dots: Locations of the instrument ports. In both positions, the optic port is located at the umbilicus. The two instrument ports are inserted at anatomic locations that enable a manipulation angle of 60°. The axis of the camera is between the axes of the working instruments.

Motion tracking

Measurements of the body movements were performed using the Flock of Birds real-time motion tracking device (Ascension Technology Corporation, Milton, Massachusetts, USA). The Flock of Birds real-time motion tracking system consists of a transmitter placed behind the participant, three sensors attached to the body, and hardware units connected to the sensors, the transmitter, and a laptop computer (Figure 2a). The sensors were attached to the head with a headband, to the skin at the level of spinous process Th1 and to the body of the sacrum S1 of the participant to track the movements.

The transmitter of the motion-tracking device creates an electromagnetic field. The motion tracker uses this electromagnetic field to determine the orientation of the sensors in relation to the x-axis, y-axis, and z-axis of the transmitter using the Euler format (roll, elevation, and azimuth) (Figure 2b). By calculating the difference between the orientation of two sensors, the angles of the cervical and thoracolumbar spine can be determined in three dimensions.

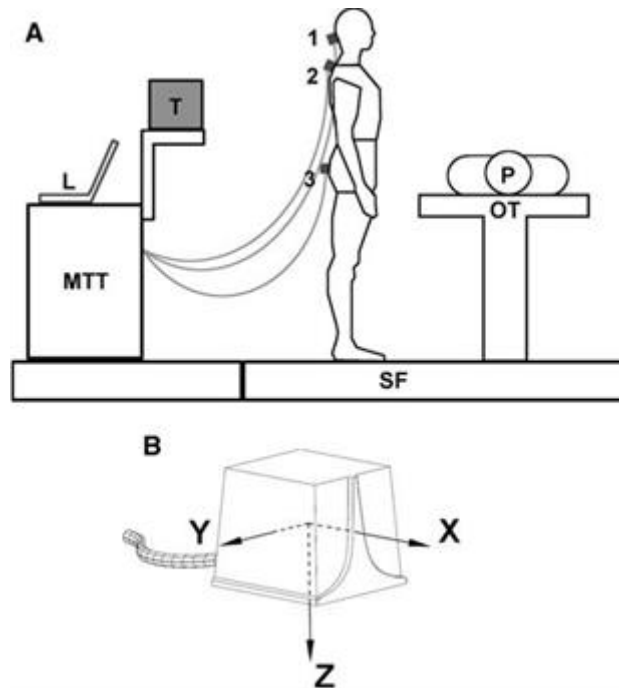


Figure 2: The motion-tracking device setup. a) Attachment of the sensors to the operating surgeon. MTT: motion-tracker trolley, T: transmitter, SF: sterile field, L: laptop, P: patient, OT: operation table. 1) sensor on the head, 2) sensor on Th1, and 3) sensor on S1. b) A sensor with projection of the axes used to calculate the body angles. Drawing courtesy of Ascension Technology Corporation. Used with permission.

Before scrubbing, the sensors were mounted to the head and body of the participating surgeon. The surgical gown could be worn over the sensors so the sterile environment was not compromised during the measurements. The motion-tracking software was configured to measure the body posture with an interval of approximately 0.33 s.

The recording was started at the introduction of the trocars and stopped at the moment of gallbladder extraction. The phases between these moments (preparation, clipping, gallbladder dissection and coagulation-suction) consist mainly of long static-posture episodes disrupted by short intervals of instrument changes when the extremities and the torso move. Research has shown that within these phases, approximately 75% of the time is spent in a static body posture.⁸ It is believed that the prolonged awkward postures during these long static-posture episodes are the main cause of neck and back problems in laparoscopic surgery.

Ergonomic principles

Postural muscles are always active while standing to counteract the forces exerted by gravity on the body mass. The activity of the muscles is minimized when the body parts are in a vertical line and the moments produced by gravity are at a minimum. The activity of the muscles around the cervical vertebral column is mainly determined by the weight and position of the head and the tension in the ligaments. Prolonged extreme forward flexion (>30 degrees) can cause complaints of the neck muscles.¹² Extension is done when looking upward. An upward gaze causes higher load on the ocular muscle¹³ and is hypothesized to enlarge the ocular surfaces leading to visual strain due to increased tear vaporisation.¹⁴ Rotation of the neck further than 35 degrees causes the muscle load to increase dramatically.¹⁵

Flexion and extension of the back is mainly facilitated in the lumbar spine.¹⁶ The activity of the associated back muscles is likewise determined by weight and position of the trunk. Flexion is initiated by the abdominal muscles and the iliopsoas and maintained by the m. erector spinae during a static posture. The abdominal muscles assist in flexion by increasing the abdominal pressure and producing an abdominal spring force, thereby reducing the work needed by the m. erector spinae

needed to maintain flexion.¹⁷ Research indicates there is an increased prevalence of low back pain in workers who have to bend or twist their back during labour hours.¹²

Ergonomic assessment

To estimate the ergonomic quality of the surgeon's posture, rotations in the thoracolumbar and cervical spines were calculated for the three anatomic planes:

- The horizontal plane (axial rotation)
- The sagittal plane (flexion/extension)
- The coronal plane (lateroflexion).

Additionally we measured the orientation of the head in the sagittal plane to qualify the extent of 'gaze-down viewing' in relation to the monitor position. The orientation of the head is the end product of the spine's posture and closely related to the position of the monitor.

For this study, the following optimal ergonomic body posture was chosen:

- Minimal axial rotation and lateroflexion in both the thoracolumbar and cervical spines
- Neutral position or slight flexion in the thoracolumbar and cervical spines
- Achievement of a "gaze-down" orientation of the head toward the operating field.

Data analysis

Neutral body posture

To calculate the angles of the vertebral column and the orientation of the head in neutral body posture, 15–25 reference measurements were recorded, with the operator instructed to stand in a neutral body posture: feet slightly apart, back and neck upright, arms alongside the body, and eyes focusing on a point at eye height on the opposite wall of the operating room. The mean angles and orientation were calculated and designated as neutral reference values for the body posture of the surgeon performing laparoscopic cholecystectomy in the French or American position.

Working body posture

A. Flexion/extension of the cervical and thoracolumbar spine

CspineF/E and TLspineF/E in each time point was calculated with the formulas:

1. $[CspineF/E]_{working\ posture} = ([Sagittal\ plane]_{head} - [Sagittal\ plane]_{Th1}) - ([Sagittal\ plane]_{head} - [Sagittal\ plane]_{Th1})_{neutral}$
2. $[TLspineF/E]_{working\ posture} = ([Sagittal\ plane]_{Th1} - [Sagittal\ plane]_{S1}) - ([Sagittal\ plane]_{Th1} - [Sagittal\ plane]_{S1})_{neutral}$

Negative values indicate flexion and positive values indicate extension.

B. Torsion of the cervical and thoracolumbar spine

NeckT and BackT in each time point was calculated with the formulas:

1. $[CspineT]_{working\ posture} = ([Transversal\ plane]_{head} - [Transversal\ plane]_{Th1}) - ([Transversal\ plane]_{head} - [Transversal\ plane]_{Th1})_{neutral}$
2. $[TLspineT]_{working\ posture} = ([Transversal\ plane]_{Th1} - [Transversal\ plane]_{S1}) - ([Transversal\ plane]_{Th1} - [Transversal\ plane]_{S1})_{neutral}$

C. Lateroflexion of the cervical and thoracolumbar spine

CspineLF and TLspineLF in each time point was calculated with the formulas:

1. $[CspineLF]_{working\ posture} = ([Frontal\ plane]_{head} - [Frontal\ plane]_{Th1}) - ([Frontal\ plane]_{head} - [Frontal\ plane]_{Th1})_{neutral}$
2. $[TLspineLF]_{working\ posture} = ([Frontal\ plane]_{Th1} - [Frontal\ plane]_{S1}) - ([Frontal\ plane]_{Th1} - [Frontal\ plane]_{S1})_{neutral}$

D. Orientation of the head in the sagittal plane

HeadOSP in each time point was calculated with the formula:

1. $[HeadOSP]_{working\ posture} = [Sagittal\ plane]_{head} - [Sagittal\ plane]_{neutral\ head}$

Statistical analyses

A Wilcoxon signed-rank test was used to compare the mean operating time of the French with the American position. The same statistical test was used to compare the body posture and the percentage of operation time within an ergonomically acceptable range. In all comparisons, a p value lower than 0.05 was considered statistically significant. To calculate the variance in the working body posture of the individual surgeons, the analysis of variance (ANOVA) formula for pooled variance was used to calculate the pooled standard deviation. The data was processed with SPSS 20.0.0.1 (SPSS, Chicago, IL, USA).

Results

Data characteristics

The mean recording time was 20.8 min per procedure and did not differ between the French and American procedures (21.6 vs 20.0 min; $p = 0.715$). No complications occurred, and all the procedures could be completed laparoscopically. All the patients were discharged from the hospital without any adverse events the day after the procedure.

Mean body angles

Table 2 shows the mean body angles for the different movement directions during the laparoscopic cholecystectomy in the French and American position. No statistically significant difference was found between the French and American position in terms of cervical spine flexion/extension ($p = 0.273$), thoracolumbar spine flexion/extension ($p = 0.273$), cervical spine torsion ($p = 0.715$), thoracolumbar spine torsion ($p = 0.465$), cervical spine lateroflexion ($p = 0.144$), or thoracolumbar spine lateroflexion ($p = 0.465$).

Table 2: Mean body angles in the sagittal, horizontal and coronal plane (values in degrees +/- SD)

Sagittal plane							
CspineF/E				TLspineF/E			
	French	American	p		French	American	p
Mean	1.9±5.6	-3.4±5.6	0.273	Mean	-5.4±4.0	-1.9±3.3	0.273
Horizontal plane							
CspineT				TLspineT			
	French	American	p		French	American	p
Mean	-0.4±6.2	-0.3±7.5	0.715	Mean	3.2±4.9	-2.9±3.9	0.465
Coronal plane							
CspineLF				TLspineLF			
	French	American	p		French	American	p
Mean	1.3±5.1	3.0±6.1	0.144	Mean	-2.2±4.6	0.7±5.1	0.465

Relative frequencies and time percentage of operation time within ergonomic acceptable range

To obtain insight into the percentage of time spent within different body angle ranges, the relative frequencies of the body angles were calculated. The relative frequency histograms of the cervical and thoracolumbar angles in the sagittal, horizontal, and coronal planes are represented in figures 3 and the head orientation is represented in Figure 4.

In the horizontal plane, no significant differences were found in the percentage of operating time within an ergonomically acceptable range in the cervical spine (French position, 97.0%; American position, 82.8%; $p = 0.144$) or in the thoracolumbar spine (French position, 94.7%; American position, 98.6%; $p = 0.144$).

Regarding the operating time within an ergonomic acceptable range in the sagittal plane, no significant difference was found in the cervical spine (French position, 71.5%; American position, 71.5%; $p = 0.273$) or in the thoracolumbar spine (French position, 97.5%; American position, 95.1%; $p = 0.715$).

In the coronal plane, no significant differences were found in the percentage of operating time within an ergonomically acceptable range in the cervical spine (French position, 98.4%; American position, 97.0%; $p = 0.715$) or in the thoracolumbar spine (French position, 98.3%; American position, 97.4%; $p = 1.000$).

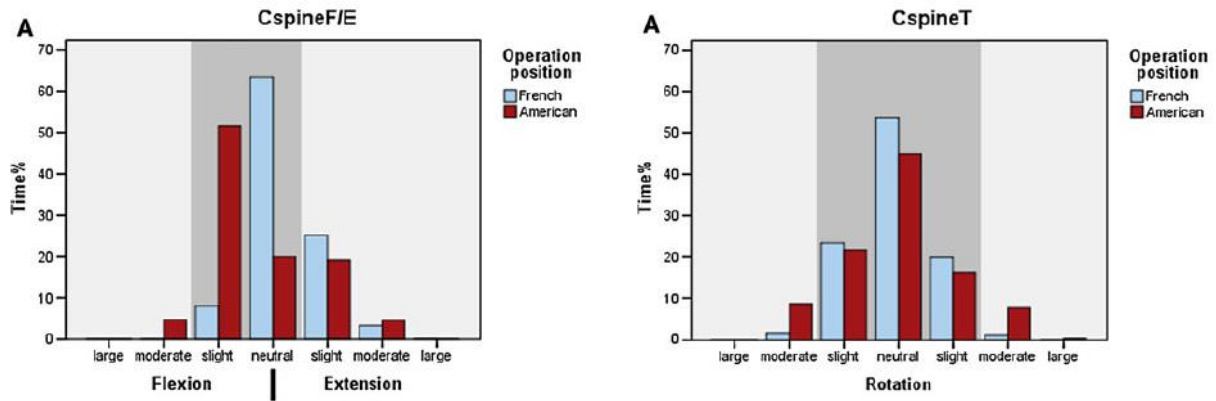


Figure 3a: Relative frequency histograms showing flexion/extension of the cervical (CspineF/E) and thoracolumbar (TLspineF/E) spine. The body angles in the sagittal plane are categorized in large flexion (lower than -35 degrees), moderate flexion (-35 to -15 degrees), slight flexion (-15 to -5 degrees), neutral position (-5 to +5 degrees), slight extension (+5 to +15 degrees), moderate extension (+15 to +35 degrees) and large extension (higher than +35 degrees). The gray coloured columns indicate the ergonomically acceptable range (-15° flexion to 5° extension).

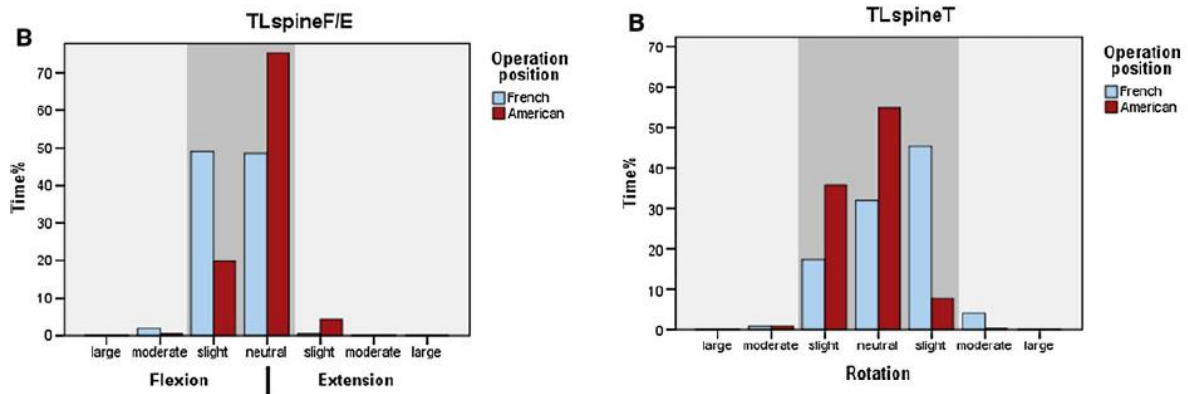


Figure 3b: Relative frequency histograms showing axial rotation in the cervical (CspineT) and thoracolumbar (TLspineT) spine. Rotation is categorized in neutral position (-5 to +5 degrees) and in slight rotation (5 to 15 degrees), moderate rotation (15 to 35 degrees) and large rotation (higher than 35 degrees). The gray coloured columns indicate the ergonomically acceptable range (<15°).

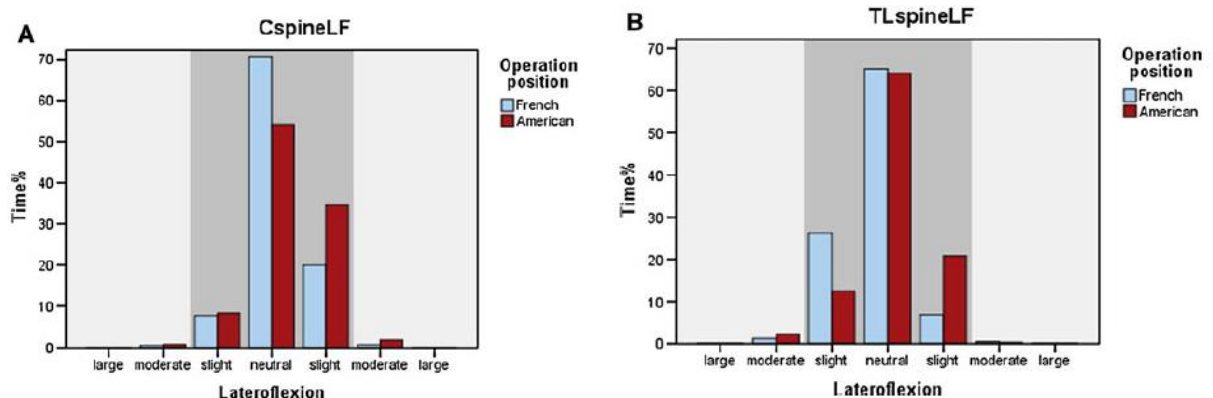


Figure 3c: Relative frequency histograms showing lateroflexion in the cervical (Cspine LF) and thoracolumbar (TLspineLF) spine. Lateroflexion is categorized in neutral position (-5 to +5 degrees) and in slight lateroflexion (5 to 15 degrees), moderate lateroflexion (15 to 35 degrees) and large lateroflexion (higher than 35 degrees). The gray coloured columns indicate the ergonomically acceptable range (<15°).

Table 3 and figure 4 show the results for the head orientation in the sagittal plane. The French and the American position did not differ in terms of the head orientation in the sagittal plane ($p = 0.465$).

Table 3: Mean head orientation in the sagittal plane

<i>HeadOSP</i>			
	French	American	p
Mean	-6.3±5.6	-6.3±5.6	0.465

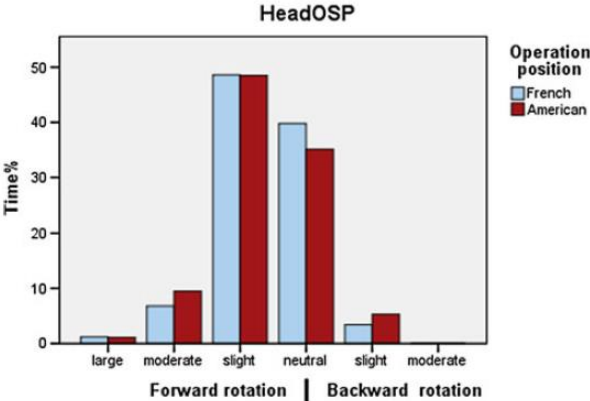


Figure 4: Relative frequency histogram showing the orientation of the head (HeadOSP). The orientation angles are categorized in large flexion (-35 to -25 degrees), moderate flexion (-25 to -15 degrees), slight flexion (-15 to -5 degrees), neutral position (-5 to +5 degrees), slight extension (+5 to +15 degrees) and moderate extension (+15 to +25 degrees).

Discussion

Laparoscopic surgery provides well-established advantages for the patient, but the operating team is confronted with ergonomic challenges. This study compared the ergonomic quality of the surgeon's body posture and the pattern of postural changes during laparoscopic cholecystectomy performed in the French and American position. To our knowledge, this was the first study to use an intraoperative motion-tracking device to perform a three-dimensional measurement of the surgeon's body posture during a laparoscopic cholecystectomy in a MIS suite.

Motion analysis of the vertebral column suggested that the surgeon's posture does not differ significantly between the French and the American position in a MIS suite. Furthermore, no statistical significant difference was found in the percentages of the time surgery was performed within an ergonomic acceptable range. In both positions, most of the time was spent within an ergonomic acceptable range. This is in contrast with results of research that assessed the ergonomics of the two operating positions in a virtual reality simulator.¹⁸ The results of this study showed better ergonomics of the vertebral column and upper extremities in the French position. A possible explanation for the discrepancy in results between this study and the current study is the adjustability of the multiple suspended monitors in the MIS suite. By adjusting the position of the monitor in the MIS suite, the surgeon's tendency to rotate the cervical and thoracolumbar spine in the American position might have been minimized to an acceptable level.

Although not statistically significant, the relative frequency histogram of cervical flexion suggests that the neck of the surgeon may be slightly more flexed for a higher percentage of the operating time in the American position (51.5 %) than in the French position (8.0 %). In the posture of the back, the contrary is found. The back is slightly more flexed for a higher percentage of the time in the French position (49.0 %) than in the American position (19.8 %). On the basis of the team positioning, we could reason that the slight thoracolumbar flexion in the French position could be caused by a greater distance between the surgeon and the operating field in the French position. This distance has to be bridged by a slight bending forward. The thoracolumbar flexion forward leads in turn to a decreased flexion of the neck in the French position compared with the American position. However, because the adaptation of the thoracolumbar spine to the work environment is within an ergonomic acceptable range (-15 degrees flexion to 5 degrees extension), the surgeon probably faces no increased risk of musculoskeletal problems.

Different variables can influence the neutral and working body postures in the operating room. For instance, in a study examining the ergonomic aspects of laparoscopic surgery, surgeons with less than 2 years experience were significantly more affected by ergonomically inefficient environments in the operation room than those with longer experience. We tried to minimize the effects of these variables in two ways: on the basis of experience (a group of two residents and two experienced surgeons were selected) and on the basis of education (one resident and one experienced surgeon were educated in the French, whereas the remaining resident and experienced surgeon were educated in the American position). Furthermore, the crossover design used in this study made it possible to correct for individual differences in working body posture between the participating surgeons. A weakness of this study and a potential hazard for type 2 errors was the small sample size.

Limitations

Some ergonomic issues could not be answered with this study. First, the relation between the surgeon's body length and body posture during surgery was not investigated. Theoretically, in the French position, the work field is further away from the surgeon. Therefore, a tall surgeon with long upper extremities can bridge this distance to the operating field easier while maintaining a straight back posture. Second, the size of the patient was not taken into account. The distance between the work field and the surgeon increases as the size of the patient increases. Therefore, a procedure on a tall patient could lead to a less comfortable posture of the vertebral column. Considering the position

of the surgeon in the operating team, this could especially be the case in the French position. Third, in this study, only the spine was taken into account. Additional in vivo measurements of the shoulder, arm, and wrist angles could provide more information about the amount of strain on the upper body in the French and American position in a MIS suite. This could be particularly interesting for the American position, in which the surgeon has the tendency to hold his upper extremities in an uncomfortable position due to the location of the instrument ports and the angle of the axes of the instruments.¹⁸ To demonstrate the importance of these factors during live operations, further studies are necessary. Nonetheless, this comparative study indicates that the posture of the vertebral column and the head orientation in the sagittal plane do not differ significantly between the French and American position in a modern MIS suite.

Conclusion

In conclusion, this comparative ergonomic study indicates that there is no significant difference between operating posture of the vertebral column in the French and American position in a modern MIS suite.

References

1. Berggren U, Gordh T, Grama D, Haglund U, Rastad J, Arvidsson D. Laparoscopic versus open cholecystectomy: hospitalization, sick leave, analgesia, and trauma responses. *Br J Surg*. 1994;81:1362–1365.
2. Hendolin HI, Paakonen ME, Alhava EM, Tarvainen R, Kempainen T, Lahtinen P. Laparoscopic or open cholecystectomy: a prospective randomised trial to compare postoperative pain, pulmonary function, and stress response. *Eur J Surg*. 2000;166:394–399.
3. Richards C, Edwards J, Culver D, Emori TG, Tolson J, Gaynes R. Does using a laparoscopic approach to cholecystectomy decrease the risk of surgical site infection? *Ann Surg*. 2003;237:358–362.
4. Schellekens PC, Bijnen AB, Honing M, Lourens J, de Ruyter P. Results of the introduction of laparoscopic cholecystectomy on morbidity and mortality of gallbladder surgery in a large regional hospital. *Ned Tijdschr Geneesk*. 1995;139:723–727.
5. Park A, Lee G, Seagull J, Meenagh N, Dexter D. Patients benefit while surgeons suffer: an impending epidemic. *J Am Coll Surg*. 2010;210:306–313.
6. Erfanian K, Luks FI, Kurkchubasche AG, Wesselhoeft CW Jr, Tracy TF Jr. In-line image projection accelerates task performance in laparoscopic appendectomy. *J Pediatr Surg*. 2003;38:1059–1062.
7. Hemal AK, Srinivas M, Charles AR. Ergonomic problems associated with laparoscopy. *J Endourol*. 2001;15:499–503.
8. Vereczkei A, Feussner H, Negele T, Fritzsche F, Seitz T, Bubbs H, Horvath OP. Ergonomic assessment of the static stress confronted by surgeons during laparoscopic cholecystectomy. *Surg Endosc*. 2004;18:1118–1122.
9. Van Det MJ, Meijerink WJHJ, Hoff C, Van Veelen MA, Pierie JPEN. Optimal ergonomics for laparoscopic surgery in minimally invasive surgery suites: a review and guidelines. *Surg Endosc*. 2009;23:1279–1285.
10. Van Veelen MA, Jakimowicz JJ, Kazemier G. Improved physical ergonomics of laparoscopic surgery. *Min Invas Ther Allied Technol*. 2004;13:161–166.
11. Mishra RK (2013) Textbook of practical laparoscopic surgery, 3rd edn. Jaypee Brothers Medical Publishers, New Delhi.
12. Ariëns GA, Bongers PM, Douwes M, Miedema MC, Hoogendoorn WE, van der Wal G, Bouter LM, van Mechelen W. Are neck flexion, neck rotation, and sitting at work risk factors for neck pain? Results of a prospective cohort study. *Occup Environ Med*. 2001;58:200–207.
13. Jaschinski W, Heuer H, Kylian H, Koitcheva V. Visual comfort at the VDU workplace and oculomotor parameters as a function of vertical inclination of gaze direction. From Experience to Innovation, IEA'97, Vol. 5. Finnish Institute of Occupational Health, Helsinki, 1997;44–46.
14. Sotoyama M, Abe S, Jonai H, Villanueva M, & Saito S. Improvement of visual comfort of VDT workers from the aspects of vertical gaze direction and tear volume. In Proceedings of the 13th Triennial Congress of the International Ergonomics Association. Tampere 1997;5:59–61.
15. Snijders CJ, Van Dijke GH, Roosch ER. A biomechanical model for the analysis of the cervical spine in static postures. *J Biomech*. 1991;24(9):783–792.
16. Farfan HF. Muscular mechanism of the lumbar spine and the position of power and efficiency. *Orthop Clin North Am*. 1975;6(1):135–144.
17. Cholewicki J, Juluru K, McGill SM. Intra-abdominal pressure mechanism for stabilizing the lumbar spine. *J Biomech*. 1999;32(1):13–17.
18. Youssef Y, Lee G, Godinez C, Sutton E, Klein RV, George IM, Seagull FJ, Park A. Laparoscopic cholecystectomy poses physical injury risk to surgeons: analysis of hand technique and standing position. *Surg Endosc*. 2011;25(7):2168–2174.

Chapter 5

DEVELOPMENT OF A STANDARDIZED TRAINING COURSE FOR LAPRASOCOPIC PROCEDURES USING DELPHI METHODOLOGY

*Martijn S. Bethlehem, Kelvin H. Kramp, Marc J. van Det, Henk O. ten Cate Hoedemaker,
Nic J.G.M. Veeger, Jean-Pierre E.N. Pierie*

Journal of Surgical Education 2014;71(6):810-816

On behalf of the expert panel representing the research group of the North-East Surgical School of the Netherlands: J.A. Apers, Medical Center Leeuwarden. G.I.J.M. Beerthuizen, Martini Hospital Groningen. R.J.I. Bosker, Deventer Hospital. E.B. van Duyn, Medisch Spectrum Twente. K. Havenga, University Medical Center Groningen. P.H.J. Hemmer, University Medical Center Groningen. Chr. Hoff, Medical Center Leeuwarden. H.S. Hofker, University Medical Center Groningen. F.W.H. Kloppenberg, Bethesda Hospital Hogeveen. S.A. Koopal, Medical Center Leeuwarden. E.A. Kouwenhoven, Twenteborg Hospital. C. Krikke, University Medical Center Groningen. E.R. Manusama, Medical Center Leeuwarden. E.J. Mulder, Antonius Hospital Sneek. V.B. Nieuwenhuijs, Isala Clinics Zwolle. E.G.J.M. Pierik, Isala Clinics Zwolle. R.A. Schasfoort, Scheper Hospital Emmen. A.P.M. Stael, Martini Hospital Groningen.

Abstract

Background: Content, evaluation and certification of laparoscopic skills and procedure training lack uniformity among different hospitals in the Netherlands. Within the process of developing a new regional laparoscopic training curriculum, a uniform and transferrable curriculum was constructed for a series of laparoscopic procedures. The aim of this study was to determine regional expert consensus regarding the key steps for laparoscopic appendectomy and cholecystectomy using a Delphi methodology.

Methods: Lists of suggested key steps for laparoscopic appendectomy and cholecystectomy were created using surgical textbooks, available guidelines and local practice. Twenty-two experts, working for teaching hospitals throughout the region, were asked to rate the suggested key steps for both procedures on a Likert scale from 1-5. Consensus was reached with Cronbach's alpha ≥ 0.90 .

Results: Out of the twenty-two experts, twenty-one completed and returned the survey (95%). Data analysis already showed consensus after the first round of Delphi on the key steps for laparoscopic appendectomy (Cronbach's alpha 0.92) and laparoscopic cholecystectomy (Cronbach's alpha 0.90). After the second round, 15 proposed key steps for laparoscopic appendectomy and 30 proposed key steps for laparoscopic cholecystectomy were rated as important (≥ 4 by at least 80% of the expert panel). These key steps were used for the further development of the training curriculum.

Conclusion: By utilizing the Delphi methodology, regional consensus was reached on the key steps for laparoscopic appendectomy and cholecystectomy. These key steps are going to be used for standardized training- and evaluation purposes in a new regional laparoscopic curriculum.

Introduction

Minimally Invasive techniques for an ever-growing number of surgical indications are adopted around the world and are becoming the gold standard for certain indications. Therefore, the need for well-trained and certified laparoscopic surgeons will increase. As working hours of surgical residents are now restricted by European directives and legislation, exposure to clinical material and the opportunity to operate is substantially limited in the current climate by comparison to twenty years ago. Therefore, a structured and focused training curriculum is needed for optimal utilization of the available training hours. The traditional “Master-Apprentice-Model” is still most commonly used to train surgical residents, sometimes in combination with pre-clinical training sessions in a skills lab. In this model, the apprentice or resident learns to perform a procedure at first by observing the master or surgeon how it needs to be done. When the resident has assisted the surgeon several times, he will gradually be allowed to perform parts of the operation under the Master’s supervision until the Apprentice can eventually perform it in total. The judgement of “proficiency” is solely based on the subjective opinion of the training surgeon. Moreover, when the resident has to learn a procedure from multiple surgeons, there will be a subsequent difference in what is taught and what is regarded as proficient. In an effort to overcome this non-transferrable and subjective method of grading performance, the Objective Structural Assessment of Technical Skills (OSATS) global rating scale has been adopted as a scoring system to evaluate a resident’s performance on both open and laparoscopic procedures.^{1,2} The OSATS global rating scale scores are saved in the digital portfolio that is implemented in all Dutch surgical training programs. A drawback of the OSATS global rating scale methodology is that it is not designed to be procedure-specific. Therefore it cannot be used for step-by-step feedback and the scoring of procedural steps. Furthermore, the OSATS global rating scale is still an instrument that displays the observer’s perception of the trainee’s technical skills that can have a certain inter-observer variability.^{3,4} Therefore content, evaluation and certification of laparoscopic skills- and procedure training lack uniformity among different hospitals in the Netherlands, but probably worldwide.

We are within the process of developing a new laparoscopic training curriculum for the North-East Surgical School of the Netherlands. We aim to construct a curriculum that provides a safe, uniform, efficient and procedure-specific training program for a series of laparoscopic procedures and make it transferrable throughout the region. Within a uniform learning curve for procedural training, we identified six different steps for each curriculum, from basic skills up to certification (Table 1). The identification was based upon the clinical and educational experience of the teaching surgeons of the surgical school of our region. Successfully completing one step will be giving access to the next step, thus only teaching the residents new skills when their own learning curve is sufficient.⁵

With the opportunity of simulating minimally invasive surgery, we aim to start training outside of the operating room. In our surgical school, the validated virtual reality simulator curriculum by SIMENDO (SIMENDO BV, Rotterdam, The Netherlands) is used to teach and assess the basic laparoscopic skills of the resident.^{6,7} Translational studies have shown that when a surgical resident successfully completes a simulator curriculum, their performance in the operating room improves.^{8,9} When successfully passing simulator practice, the resident will then learn basic laparoscopic skills at obligatory cadaver practice. What is new in our curriculum, and what distinguishes it from other existing curricula, is that we then move on to practising procedure specific skills on animal models or human cadavers. We will be using instruction videos to demonstrate the key steps while the resident performs them. We have already shown that INtraoperative Video-Enhanced Surgical procedure Training (INVEST) has a positive effect on the completion of the early learning curve for surgical procedural training by both increased efficiency and increased effectiveness.^{10,11} After this step is passed, the resident will go to the operating theatre to actually perform laparoscopic procedures on patients while being trained with the INVEST videos and supervised by an experienced instructor. The INVEST videos will be shown on one of the two (or

three) monitors available during the operations on patients, meaning a short break in actual operating. In the mean while the resident and supervisor keep complete control of the operation field, because they're being able to see the live camera feed on the other monitor(s). We have also already shown that total procedure time was not lengthened by INVEST.⁷

Table 1 **The six steps of the new laparoscopic training curriculum**

Step 1	eye-hand coordination on a simulator
Step 2	basic laparoscopic skills and safety measures in the skills lab
Step 3	specific procedural training in skills lab
Step 4	video-assisted side-by-side training in the hospital operating room
Step 5	operating under supervision in the hospital operating room
Step 6	feedback through registration of results and certification

The aim of this study was to determine expert consensus regarding the key steps required for teaching a laparoscopic appendectomy and cholecystectomy using a Delphi methodology. The outcome of the Delphi panel will be the key steps that are going to be used for creating the INVEST videos for both procedures.

By teaching all surgical residents the same key steps for every laparoscopic procedure, we aim that eventually a procedure specific assessment tool can be validated. The final goal would be to create an objective assessment, which leads to procedure specific accreditation to be given valid for every (teaching) hospital the surgical resident will be working at. There are procedure specific evaluation tools that have already been validated and are being used in clinical practise like the Global Operative Assessment of Laparoscopic Skills (GOALS) or Operative Performance Rating System (OPRS).^{12,13,14} However these tools are still used to evaluate residents who underwent non-standardized training. Evaluating surgical residents on the performance of the key steps that have been the foundation of their training curriculum is a method that, for as far as we know, has not been validated.

Methods

Study design

In order to reach consensus on the key procedural steps for teaching the laparoscopic appendectomy and cholecystectomy, the Delphi methodology was used. The Delphi method is a well-established, completely anonymous, group process in which ideas are expressed to the participants in the form of a questionnaire.^{15,16} Responses to the items in the questionnaire are collected and analysed along with added comments of the experts. This leads to adding, revising or dropping of items to be used in a second or further subsequent round until group consensus is reached.^{16,17} The Delphi method avoids the possibility that the highest positioned expert is the most influential in reaching consensus and secondly, prevents that an expert will adjust to the group opinion regardless of the evidence that supports his own opinion.

Expert panel

In the literature, there is no guideline for the number of experts required for a Delphi survey. For this study, twenty-two experts were asked to participate in the study. All were experienced and currently practicing laparoscopic surgeons who are involved in training laparoscopic procedures for residents and fellow surgeons. Furthermore, they were members of the North-East Surgical School of the Netherlands and therefore representatives from every teaching hospital and some non-teaching hospitals throughout the region. The individual experts were not informed about their fellow participants in the panel.

The Delphi questionnaire

We constructed a list of the possible key steps required to perform a laparoscopic appendectomy and cholecystectomy and they were mailed to the experts. The non-responders received digital versions as reminders. The key steps were compiled from surgical textbooks and current guidelines from the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES)^{18,19}, the European Association for Endoscopic Surgery (EAES)²⁰ and the Association of Surgeons of the Netherlands (NVvH).²¹ Each possible key step identified from these sources was included in the survey for completeness (Tables 2 and 3).

For the first round of the Delphi questionnaire each expert was asked to rate the key steps on a Likert-scale from 1 ("strongly disagree") to 5 ("strongly agree") to what extent, they believed, a step should be considered a key step and should be included in the final training curriculum. In addition, the experts were offered the opportunity to comment on each key step or clarify their ratings. This led to removing key steps because there was consensus on these key steps not being useful (> 80% of the expert panel rating it ≤ 2 after the first round). Key steps that were rated as important (≥ 4 by at least 80% of the expert panel) are going to be used for the further development of the training curriculum.

For the second round of the Delphi questionnaire, we used the comments provided by the panellists on the suggested items as input for modification of the key steps that didn't meet the above mentioned criteria (marked with an * in Tables 2 and 3). This led to the fusion of key steps or revising a key step into a more general key step. With these alterations, we are leaving more room for performing a part of the operation depending on anatomical or other situational variations. We provided additional information to clarify these key steps in an open forum discussion and gained a new opinion of the experts.

Table 2 The list of proposed key steps identified for laparoscopic appendectomy

Preoperative preparing

- Checking of instruments, devices and optics
- Positioning of the patient (right arm out, left arm alongside the patient)
- Positioning of the operating team
- Positioning of the monitors
- Placement of a gastric tube
- Antibiotic prophylaxis
- Disinfection and draping (from nipple line to os pubis)

Access and port insertion

- Open introduction using Hasson technique (SU)
- Creating pneumoperitoneum using a Veress needle
- Placing of two additional ports under direct vision (SP and LLQ)

Diagnostic laparoscopy

- Inspecting the intraperitoneal organs
- Identifying an appendix sana or appendicitis

Exposure

- Placing the patient in Trendelenburg position and tilted to the left
- Grasping the mesoappendix with the clamp through the SP port
- Retracting the appendix in the direction of the ventral abdominal wall

Taking care of the mesoappendix

- Preparation of the mesoappendix and appendicular artery*
- Placing two clips on the appendicular artery at the cecal base*
- Cutting the appendicular artery and mesoappendix*

Looping and cutting

- Placing two loops around the appendix
- Cutting the appendix between the loops

Ending the operation

- Introducing the extraction bag through the LLQ port*
- Placing the appendix in the extraction bag*
- Irrigation and suction around the appendicular stump on indication
- Removing the appendix
- Removing the ports under direct vision*
- Closing of fascial defects > 5mm*
- Closing of the skin with intracutaneous sutures
- Removing the gastric tube

SU = subumbilical; SP = suprapubic; LLQ = left lower quadrant. * key steps discussed in round 2.

Table 3 **The list of proposed key steps identified for laparoscopic cholecystectomy**

Preoperative preparing

- Checking of instruments, devices and optics
- Positioning of the patient (right arm alongside of the patient)*
- Positioning of the operating team*
- Positioning of the monitors
- Placement of a gastric tube
- No indication for antibiotic prophylaxis
- Disinfection and draping (from nipple line to well below the umbilicus)

Access and port insertion

- Open introduction using Hasson technique (SU)
- Creating pneumoperitoneum using a Veress needle
- Placing of three additional ports under direct vision (EG and 2x RUQ)

Diagnostic laparoscopy

- Inspecting the intraperitoneal organs*

Exposure

- Placing the patient in reversed Trendelenburg position and tilted to the left
- Retracting the fundus from the most lateral port in a cephalad and anterior direction
- Adhesiolysis flush on the gallbladder
- Identifying the infundibulum and the hepatoduodenal ligament
- Retracting the infundibulum in a caudal and lateral direction

Opening the peritoneum

- Opening the peritoneal envelope from the infundibulum
- Opening the peritoneum medial and lateral from the infundibulum to the fundus

Dissection of the triangle of Callot

- Dissection of fat and fibrous tissue step by step and flush on the gallbladder
- Exposing the cystic duct at the gallbladder
- Identifying the cystic duct
- Exposing the cystic artery at the gallbladder
- Identifying the cystic artery

Critical View of Safety

- Establishing the Critical View of Safety
- Documenting the Critical View of Safety

Intraoperative cholangiography

- Placing a clip on the cystic duct where it enters the gallbladder
- Cutting the cystic duct until gall is discharged
- Catheterising with flushed cholangi catheter and occluding the cystic duct around it
- Creating and interpreting the intraoperative cholangiography

Clipping and cutting

- Clipping the cystic artery (two clips central and one at the side of the gallbladder)
- Cutting the cystic artery
- Clipping the cystic duct (two clips central and one at the side of the gallbladder)
- Cutting the cystic duct

Retrograde cholecystectomy

- Further opening the peritoneum
- Dissecting the gallbladder from the liver bed
- Establishing haemostasis of the liver bed

Ending the operation

- Introducing the extraction bag through the SU port*
- Placing the gallbladder in the extraction bag and removing it through the SU port*
- Removing the ports under direct vision
- Closing of fascial defects > 5mm*
- Closing of the skin with intracutaneous sutures
- Removing the gastric tube

EG = epigastric; RUQ = right upper quadrant; SU = subumbilical. * key steps discussed in round 2.

Statistical analysis and consensus

Cronbach's α was chosen as the statistical index to quantify the reliability of the group of panellists.¹⁶ When the responses of the experts are highly correlated, in this study when Cronbach's $\alpha > 0.90$, they are considered as internally consistent and thus suggesting consensus. Means and standard deviations were calculated for all key steps. Cronbach's α was calculated for laparoscopic appendectomy and laparoscopic cholecystectomy. All statistical analysis was performed using SAS version 9.2.

Results

Of the twenty-two experts asked to participate in the Delphi panel, twenty-one (95%) completed and returned the survey. Data analysis of the first round already showed consensus on the key steps for laparoscopic appendectomy (Cronbach's alpha 0.92) and laparoscopic cholecystectomy (Cronbach's alpha 0.90). After the second round 15 key steps for the laparoscopic appendectomy and 30 key steps for the laparoscopic cholecystectomy were rated as important (Tables 4 and 5). These key steps are going to be used for the further development of the training curriculum.

Table 4 **The key steps for laparoscopic appendectomy**

Preoperative preparing

- Positioning of the patient (right arm out, left arm alongside the patient)
- Positioning of the monitors
- Disinfection and draping (from nipple line to os pubis)

Access and port insertion

- Open introduction using Hasson technique (SU)
- Placing of two additional ports under direct vision (SP and LLQ)

Diagnostic laparoscopy

- Inspecting the intraperitoneal organs
- Identifying an appendix sana or appendicitis

Exposure

- Placing the patient in Trendelenburg position and tilted to the left
- Retracting the appendix in the direction of the ventral abdominal wall

Taking care of the mesoappendix

- Clipping and cutting or coagulating the appendicular artery with diathermia depending on anatomy

Looping and cutting

- Placing two loops around the appendix
- Cutting the appendix between the loops

Ending the operation

- Protecting the abdominal wall against contamination by removing the appendix in an extraction bag or in the trocar depending on the situation
 - Removing the ports under direct vision
 - Closing of fascial defects > 5mm
-

SU = subumbilical; SP = suprapubic; LLQ = left lower quadrant.

Table 5 The key steps for laparoscopic cholecystectomy

Preoperative preparing

- Positioning of the patient (right arm alongside of the patient)
- Positioning of the operating team
- Disinfection and draping (from nipple line to well below the umbilicus)

Access and port insertion

- Open introduction using Hasson technique (SU)
- Placing of three additional ports under direct vision (EG and 2x RUQ)

Diagnostic laparoscopy

- Inspecting the intraperitoneal upper abdominal organs

Exposure

- Placing the patient in reversed Trendelenburg position and tilted to the left
- Retracting the fundus from the most lateral port in a cephalad and anterior direction
- Adhesiolysis flush on the gallbladder
- Identifying the infundibulum and the hepatoduodenal ligament
- Retracting the infundibulum in a caudal and lateral direction

Opening the peritoneum

- Opening the peritoneal envelope from the infundibulum
- Opening the peritoneum medial and lateral from the infundibulum to the fundus

Dissection of the triangle of Callot

- Dissection of fat and fibrous tissue step by step and flush on the gallbladder
- Exposing the cystic duct at the gallbladder
- Identifying the cystic duct
- Exposing the cystic artery at the gallbladder
- Identifying the cystic artery

Critical View of Safety

- Establishing the Critical View of Safety
- Documenting the Critical View of Safety

Clipping and cutting

- Clipping the cystic artery (two clips central and one at the side of the gallbladder)
- Cutting the cystic artery
- Clipping the cystic duct (two clips central and one at the side of the gallbladder)
- Cutting the cystic duct

Retrograde cholecystectomy

- Further opening the peritoneum
- Dissecting the gallbladder from the liver bed
- Establishing haemostasis of the liver bed

Ending the operation

- Protecting the abdominal wall against contamination by removing the gallbladder in an extraction bag depending on the situation
 - Removing the ports under direct vision
 - Closing of fascial defects > 5mm
-

EG = epigastric; RUQ = right upper quadrant; SU = subumbilical.

Discussion

The purpose of this study was to compile a list of key steps for the creation of INVEST instructional videos for laparoscopic appendectomy and cholecystectomy. The final lists were developed through a survey using the Delphi methodology. They represent consensus of experts in training minimally invasive surgery from the North-East Surgical School of the Netherlands. This is a next step in the development of a new standardized training course for laparoscopic procedures. The procedural steps in laparoscopy cholecystectomy and appendectomy that have been published in earlier research have been determined and evaluated by a relatively small group of experts.^{22,23} To our knowledge, this is the first study that uses a previously validated method in combination with a large group of twenty-one participating experts to establish consensus on which specific procedural steps should be seen as key steps for a standard laparoscopic procedure.^{15,16}

The most important point of attention is that the identified key steps can only be used for treating uncomplicated appendicitis or gallbladder disease. For example, when performing an appendectomy for retrocaecal appendicitis, the key steps don't include the then needed mobilization of the right colon. We think that the traditional Master-Apprentice-Model is momentarily the most frequent used method to learn to deal with this specific situation. The same applies for dealing with a necrotic appendicular stump, an abscess, the decision to drain or not to drain and the indications and timing for a decision to convert to an open procedure. Similar situations that are not covered with the key steps can also be encountered when performing a cholecystectomy. For example, dealing with an intra-operative perforation of the gallbladder, with or without spillage of stones, or an acute cholecystitis. The implementation of teaching procedural decision making should be during (procedure specific) training in the skills lab. Studies using a Cognitive Task Analysis to identify the key decision making-points, potential errors and complications, and problem solving strategies seem to be valuable to design a method to teach these non-technical aspects of operative performance.^{24,25} Studies that translate the transfer of these skills to the operating room have not yet been performed.

Consensus for both procedures was already achieved with the first round of the Delphi questionnaire. Still, for some of the more important key steps of both procedures we didn't reach >80% of the expert panel to rate them as important. Analysis of the comments from the panellists led to rephrasing some of the key steps. These slightly altered key steps were presented to the expert panel and approved in an open forum discussion. We used this method for the second Delphi round, because some of the key steps in the first round were not unequivocally formulated.

For both the laparoscopic appendectomy and cholecystectomy, the first round of the Delphi questionnaire showed three major points of discussion. First and most notable was the difference between the need to use laparoscopic equipment on trolleys or having equipment available in columns attached to a ceiling-mounted suspension system. The latter mostly being available in modern(ized) operating rooms designed as dedicated minimally invasive surgery (MIS) suites. In most hospitals in our teaching region both situations do occur, so we needed to combine key steps for the preoperative preparation to suit both needs. We reached consensus for both procedures on positioning the patient in such manner that an equipment trolley can be set up on the floor and still optimising efficient and ergonomic use by the operating team.

Secondly, the method of extraction of the appendix or gallbladder proved to be much dependent on the preference of the surgeon, e.g. through which trocar opening, whether or not to use an extraction bag and if this depends on the degree of contamination. These factors are most of the time not predictable before actually performing the laparoscopy. By making these factors variable within the revised key step, we reached consensus in the second Delphi round.

Closing the fascia of the trocar sites >5mm after laparoscopy was a third point of discussion for both operations. Six experts (28%) responded that closure of the trocar sites can be difficult, mostly when the patient has more subcutaneous fat, and that they don't want to make bigger wounds to close the fascia at all costs. Our intention with this key step was to teach closure of the fascia to minimize the incidence of trocar site hernias. When we explained this to the experts, who didn't favour this key step, they agreed that the intention of closing bigger fascia defects is a key stone of laparoscopic surgery.

For the laparoscopic cholecystectomy the expert panel was much divided on whether or not to perform a routine intraoperative cholangiography (IOC) in the training for surgical residents. Therefore, we went back to the opinion of the NVvH reflected in their latest guideline. They advice that, although IOC has a high sensitivity and specificity for detecting choledocholithiasis, best practice is to diagnose and treat choledocholithiasis preoperatively.²¹ We are also taking into account that IOC lengthens the procedure and has its own morbidity.

Conclusion

The Delphi methodology was successfully used to determine consensus regarding the operative key steps for laparoscopic appendectomy and cholecystectomy. These key steps are going to be used for creating procedure specific instruction videos as a next step towards standardized procedural training in a new regional laparoscopic training curriculum for the North-East Surgical School of the Netherlands. By using the Delphi methodology we hope to reach a high level of participation when these key steps are implemented in the assessment of standard laparoscopic procedures.

References

- 1 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via innovative “bench station” examination. *Am J Surg.* 1997;173:226-230.
- 2 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchinson C, *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84:273-278.
- 3 van Hove PD, Tuijthof GJM, Verdaasdonk EG, Stassen LP, Dankelman J. Objective assessment of surgical skills. *Br J Surg.* 2010;97:972-987.
- 4 Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. *Surg Endosc.* 2011;25:356-366.
- 5 van Det MJ. Training, efficiency and ergonomics in minimal invasive surgery (Thesis). Groningen: University of Groningen. 2012;92-94.
- 6 Verdaasdonk EG, Stassen LP, Monteny LJ, Dankelman J. Validation of a new basic virtual reality simulator for training of basic endoscopic skills: the SIMENDO. *Surg Endosc.* 2006;20:511-518.
- 7 Verdaasdonk EG, Stassen LP, Schijven MP, Dankelman J. Construct validity and assessment of the learning curve for the SIMENDO endoscopic simulator. *Surg Endosc.* 2007;21:406-412.
- 8 Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room – a randomized controlled trial. *Am J Surg.* 2010;199:115-120.
- 9 Seymour NE, Gallagher AG, Roman SA, O’Brien MK, Bansal VK, Andersen DK, Satava RM. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg.* 2002;236:458-464.
- 10 van Det MJ, Meijerink WJHJ, Hoff C, Middel LJ, Koopal SA, Pierie JPEN. The learning effect of intraoperative video-enhanced surgical procedure training. *Surg Endosc.* 2011;25:2261-2267.
- 11 van Det MJ, Meijerink WJHJ, Hoff C, Middel B, Pierie JPEN. Effective and efficient learning in the operating theater with intraoperative video-enhanced surgical procedure training. *Surg Endosc.* 2013;27:2947-2954.
- 12 Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107-113.
- 13 Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg.* 2007;204:308-313.
- 14 Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery.* 2005;138:640-649.
- 15 Dalkey NC. The Delphi method: an experimental study of group opinion. Santa Monica,CA: RAND Corporation; 1969.
- 16 Graham B, Regehr G, Wright JG. Delphi as a method to establish consensus for diagnostic criteria. *J Clin Epidemiol.* 2003;56:1150-1156.
- 17 Palter VN, MacRae HM, Grantcharov TP. Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. *Am J Surg.* 2011;201:251-259.
- 18 Guidelines for Laparoscopic Appendectomy. 2009. Website Society of American Gastrointestinal and Endoscopic Surgeons. Available from: <http://www.sages.org/publications/guidelines/guidelines.php>.
- 19 Guidelines for the Clinical Application of Laparoscopic Biliary Tract Surgery. 2010. Website Society of American Gastrointestinal and Endoscopic Surgeons. Available from: <http://www.sages.org/publications/guidelines/guidelines.php>.
- 20 Neugebauer EAM, Sauerland S, Fingerhut A, Millat B, Buess G. EAES Guidelines for Endoscopic Surgery. Berlin: Springer. 2006;p265-281, p291-295.

- 21 Guideline Galsteen. 2007. Website Association of Surgeons of the Netherlands. Available from: <http://www.heelkunde.nl/kwaliteit/richtlijnen/richtlijnen-definitief>.
- 22 Eubanks TR, Clements RH, Pohl D, Williams N, Schaad DC, Horgan S, Pellegrini C. An objective scoring system for laparoscopic cholecystectomy. *J Am Coll Surg.* 1999;189:566-574.
- 23 Sarker SK, Chang A, Vincent C, Darzi SA. Development of assessing generic and specific technical skills in laparoscopic surgery. *Am J Surg.* 2006;191:238-244.
- 24 Sullivan ME, Ortega A, Wasserberg N, Kaufman H, Nyquist J, Clark R. Assessing the teaching of procedural skills: can cognitive task analysis add to our traditional teaching methods? *Am J Surg.* 2008;195:20-23.
- 25 DaRosa D, Rogers DA, Williams RG, Hauge LS, Sherman H, Murayama K, Nagle A, Dunnington GL. Impact of a structured skills laboratory curriculum on surgery residents' intraoperative decision-making and technical skills. *Acad Med.* 2008;83:S68-S71.

PART III: ASSESSMENT



Chapter 6

ESTIMATING THE INTER-RATER RELIABILITY OF SURGICAL SKILLS ASSESSMENT

Kelvin H. Kramp, Marc J. van Det, Jean-Pierre E.N. Pierie

Submitted

Abstract

The interest in the reliability of surgical skills assessment has increased substantially over the past decades. Inter-rater reliability, a subform of reliability, is defined as the amount of agreement between human raters using the same assessment instrument. We discuss important aspects of the statistics and study design in the context of subjective assessment in surgical education. The aim of this paper is to equip the surgeon scientist with the statistical methods and study designs for evaluating the inter-rater reliability of surgical skills assessment and to provide designers of surgical training programs and clinical supervisors with the necessary knowledge for assessing the quality of these studies.

1. Background

The majority of current surgeons were trained according to the master-apprentice model in which a master surgeon decides whether a trainee showed sufficient improvement based on his/her own perception of the necessary skills and knowledge for surgery. However, pressure from the public and governmental institutions has led to the development of more objective assessment methods in the last decennia.^{1,2} These surgical assessment methods force clinical supervisors to quantify the quality of the observed skills on a specific set of domains relevant to the development of surgical competency. The numerical ratings can be used to monitor progression during a training program, identify strengths and weaknesses in trainees, compare the efficacy of different training curriculums, measure retention of skills after a training program and facilitate licensing in the independent treatment of uncomplicated disease.

Multiple studies have shown that a proportion of these methods are valid tools for some of these purposes.³⁻⁸ Unfortunately, a concern repeatedly addressed in these studies is the insufficient amount of agreement between raters rating the same performance, a concept also known as inter-rater reliability.¹ Inadequate reliability can impede implementation of an assessment method, because the outcome of an assessment can only be utilized if the precision is of an acceptable level. While the introduction of simulators created an opportunity for more objective and reliable assessment of psychomotor skills, the assessment of higher levels of cognitive abilities still remains a task of experienced surgeons charged with the responsibility to safely guide trainees during the acquisition of surgical skills in the highly dynamic environment of the OR. New assessment methods are continually being developed and improved to increase the inter-rater reliability of assessment by supervising surgeons. However, although reliability coefficients, which are used to measure inter-rater reliability, are one of the most important aspects of assessment methodology, those involved in surgical education are frequently unfamiliar with the rationale behind the statistics and study designs necessary for the execution of reliability research of an acceptable quality. Also, the surgeon scientist, eager to conduct research according to the highest scientific principles, can be confronted with the difficulties of choosing the right combination of statistics and methodology to estimate the validity and inter-rater reliability for a study focused on subjective assessment methodology. This can pose a problem in the field of surgical education, because understanding of the calculation methods of the intra-class correlation coefficient (ICC), the most often used statistic for calculating inter-rater reliability, is imperative for correct execution and interpretation of studies addressing the inter-rater reliability of surgical skills assessment. Previous reviews published in the surgical literature that addressed inter-rater reliability were limited in promoting a deeper understanding of inter-rater reliability.^{9,10} Therefore, the aims of this paper are:

- 1) To provide an introduction to the use and rationale of the ICC.
- 2) To discuss important aspects of study design in the calculation and evaluation of inter-rater reliability.

2.The ICC

2.1 Rationale behind the ICC

Of 52 studies included in two systematic reviews addressing inter-rater reliability of subjective assessment in surgical education, 22 studies used the ICC, 13 studies used Cronbach alpha, 3 studies used a Pearson or Spearman correlation coefficient, 3 studies used Generizability coefficient and 11 studies used various other methods.^{1,2} The ICC can therefore be considered as the most frequently used measure of inter-rater reliability in surgical education.

There are advantages of choosing the ICC to calculate inter-rater reliability instead of other correlation coefficients. The disadvantage of using the Pearson correlation coefficient for inter-rater reliability is that it only estimates the degree of association between two variables and says little about the amount of agreement between measurements. The Pearson correlation coefficient, however, continues to be used by some as a measure of inter-rater reliability until recently²⁷⁻³⁰, while it can better be reserved for the quantification of an association between two measurements that do not share metric or variance (e.g. BMI and daily caloric intake). To estimate the correlation between measurements that have the same unity, or belong to the same 'class', such as measurements performed by different raters on the same scale, the ICC is a better candidate.³ Moreover, the ICC also has the advantage of being able to measure the correlation between more than two series of measurements (e.g. ratings from 3 or more raters), while the Pearson and Spearman correlation coefficients are limited to the use for two variables.

Generizability theory is an upcoming theory for estimating reliability in the field of educational psychology. Generizability theory is based on the estimation of variance components. It gives researchers the opportunity to calculate the exact percentage of error variance each source of error is responsible for. Although these opportunities make generizability theory an attractive model, the calculation method used for estimation of variance components used in generizability varies between and within software packages (e.g. ANOVA, maximum likelihood and minimum norm quadratic unbiased estimation) while the ICC models are only based on ANOVA calculations. Moreover, some of the software packages such as SPSS and SAS are unable to cope with missing data when calculating variance components, which is a relatively common phenomenon in educational research. And third, calculation of the reliability of the assessment of a single rater (similar as the single measures ICC models), which is the measure of interest in the majority of validation studies, requires the execution of additional so-called Decision-studies, while the ICC calculations with standard software packages directly provide estimates for the use of single and average ratings.

The numerical value of the ICC can be calculated with different models. To get a general idea of what is measured with these models, the reliability coefficient calculated with the ICC can be simplified to:

$$\text{Reliability coefficient} = \text{True variance} / [\text{True variance} + \text{Error variance}] \quad 1)$$

Thus, the reliability coefficient calculated with the ICC is in essence the proportion of variance in the sample attributable to true variance. True variance is an abstract concept, but it can be estimated by subtracting the error variance from the total variance. The formula for the reliability coefficient would then become:

$$\text{Reliability coefficient} = [\text{Total variance} - \text{Error variance}] / [\text{Total variance}] \quad 2)$$

The resulting reliability coefficient is a number between 0 and 1, whereby 0 means no agreement between measurements and 1 means total agreement between measurements. The exact formulas for calculating the ICC can be found in the publication of Shrout&Fleiss.¹¹

2.2 How to choose the right model to calculate the ICC

In 2 systematic reviews addressing the validity and reliability of surgical assessment, 17 out of the 22 studies that used the ICC to calculate a reliability coefficient did not report the used calculation model.^{1,2} However, the inter-rater reliability can vary significantly depending on the model used to calculate the reliability coefficient. Lahey et al. have reported examples of 20-fold differences in size while using the same data.¹² It is therefore important to choose the right model according to the design of the study and to report which model has been used so the appropriateness of the applied ICC model can be evaluated as a part of quality assessment.

In total there are 6 different formulas to calculate the ICC: ICC-1 to ICC-3, which have different assumptions concerning raters and subjects, and type 1 or type k, which indicate a single or average measures ICC. A flowchart for choosing a model and examples of application of these models in the research field of surgical education are provided in figure 1.

Model 1 (one-way random) is suitable when the same subjects are rated by different raters during the study. This calculation model assumes subjects are not consistently rated by the same set of raters in the research setting. The calculation model allows generalization to other raters and subjects with the same characteristics, but it does not enable the users of the assessment instrument to mathematically infer the specific part of error variance that can be attributed to the variance between raters in the resulting reliability coefficient. Model 2 (two-way random), referred to as absolute agreement model, can be applied when the same subjects are all rated by the same set of raters during the study. For model 3 (two-way mixed), referred to as consistency agreement model, the same is true as for model 2, except that in this model the raters are assumed to be the exact same raters that will conduct assessment in the future. This model can only be used in the case that the included raters will be the only raters that will perform assessments in the future or the researcher is not interested in the absolute differences between ratings, but only in the inconsistencies between ratings.

The 3 models can be used to calculate the reliability for 2 types of measurements: single measures (type 1) and average measures (type k). Mathematically the type 1 coefficients are a derivative of type k coefficients. The average measures ICC will always give a higher estimate than the single measures ICC, as average measures are more reliable than single measures, however, the average measures ICC can only be used in special cases. If a researcher has used the average of a series of measurements to calculate a mean outcome to estimate reliability, and secondly, will also apply the same protocol in the future, the average measures ICC is of interest. If one of these criteria is not met, the single measures ICC is of interest. Interestingly, the average measures ICC of model 3 (ICC-3,k) is mathematically equal to Cronbach alpha.¹³

It is important to note that, that the 'subjects' do not per se have to consist of different persons. For instance, the ratings of one subject rated during different levels of experience can also be treated as multiple subjects in the calculation model.

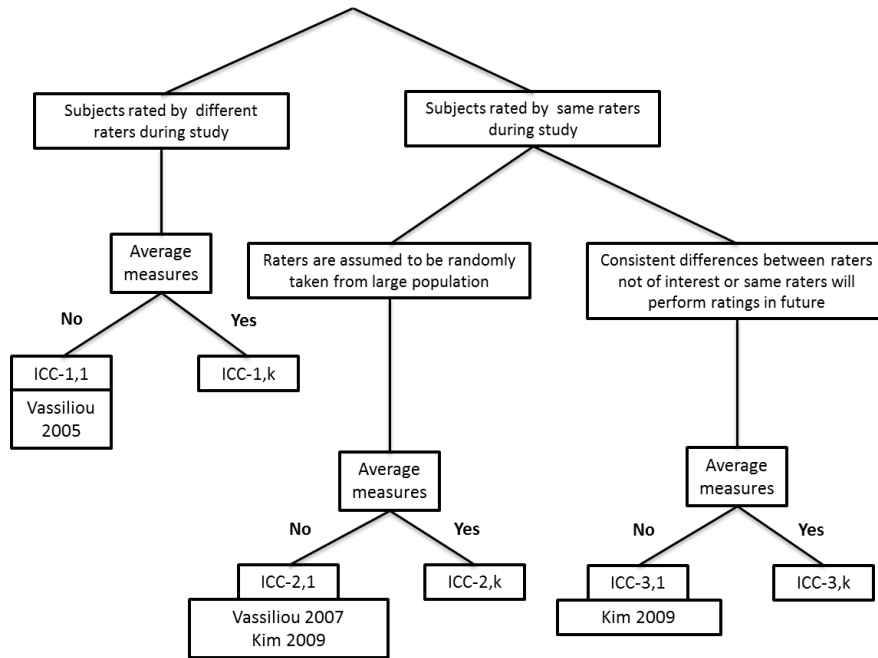


Figure 1: Flowchart for choosing a model based on study design and examples from the literature.

2.3 Calculating a sample size

Clinical supervisors can only invest a limited amount of time because of clinical burdens and there are typically only a limited number of consultant surgeons available or willing to participate. It is therefore likely that, as in most studies in medical research, the smaller the sample size the higher the feasibility of the study will be. To ensure that the sample of subjects to be rated is sufficient to achieve statistical significant results, a sample size calculation can be conducted with complex calculations published by Donner&Eliasziv or in non-integral values with the more simple, but less exact, formula published by Walter et al.^{14,15}:

$$k = 1 + (2 (2.4865)^2 N / (\ln C_0)^2 (n - 1)) \quad 3)$$

where, k = number of needed subjects at $\alpha = 0.05$ (significance level, type I error or false positive rate) and $\beta = 0.20$ (power, type II error or false negative rate), N = number of raters and C_0 is given by:

$$C_0 = (1 + (N (ICC_0 / (1 - ICC_0)))) / (1 + (N (ICC_1 / (1 - ICC_1)))) \quad 4)$$

where, ICC_0 = ICC of the null hypothesis (H_0) and ICC_1 = ICC of the alternative hypothesis (H_1). In the case of $N = 2$ there are small adjustments in the formula for the sample size calculation.¹⁴ Table 1 shows the calculated sample sizes for different values of N for $H_0:ICC = 0$ and for $H_1:ICC = 0.4$ or 0.8 . Note that, as is true for correlations in general, smaller differences between H_0 and H_1 require larger sample sizes and that small sample sizes require higher ICCs to reach statistical significance.

The sample size calculation with the formula of Walter et al. is based on ICC-1. Because this model results in the smallest ICC, it can also be used as a practical method to estimate the sample size for ICC-2 and 3. Furthermore, sample sizes are the same for the single and average measures type models.

Table 2: Sample sizes calculated with formula 3 and rounded to integral values for ICC₀ = 0 and ICC₁ values of 0.4 and 0.8 at α = 0.05 and β = 0.20.¹⁴

Numer of raters	Sample size	
	ICC ₁	
	0.4	0.8
3	16	4
4	11	3
5	8	3
10	4	2

2.4 Interpretation of the size of the ICC

The reliability coefficient indicates the proportion of the variance that can be attributed to true variance. The remaining proportion of variance can be caused by rater error, random error and/or other sources of error. If the reliability coefficient is very high or low, it is less difficult to draw conclusions than when the reliability coefficient is in between extreme values. Cut-off values used for classification of the reliability coefficient can be helpful, but are always arbitrary in nature and should be adjusted to the purpose of the measurement instrument. For formative assessment (feedback during learning), the interpretation values may be less stringent than for summative assessment (high stakes examination).¹⁶ In surgical and medical literature, a cut-off value of 0.8 has widely been adopted as the threshold for high stakes examination^{1,9,17-20}, although there is no high level evidence that supports a rationale for this specific value.²¹

Another option for interpreting the reliability coefficient, is to calculate the standard error of measurement (SEM). The SEM can be used to assess the corresponding probability distribution of the obtained score of a subject in the case that consecutive assessments would be conducted on the same subject by other raters (assuming these raters have similar characteristics as those included in the original research). The SEM can be calculated with the formula²²:

$$SEM = Sd \times \sqrt{1-ICC}$$

5)

Where Sd = the standard deviation of scores calculated for a set of ratings on the performance level of the subject of interest. This method allows a more exact interpretation of results. The SEM can be used to assess the 95% confidence interval of a single rating of a subject, assuming the rater and subject have the same characteristics as those that participated in the reliability study. Let's take the example of an assessment score with the OSATS of 11/35 of a novice (35 is the maximum score of the OSATS). If in previous studies it has been shown that the Sd for novices is 3 and the inter-rater reliability of the OSATS is 0.58, the SEM would be 1.94 and the corresponding 95% confidence interval for the assessment repeated by other raters would be 7/35 – 15/35 in rounded scores.

2.5 P-values of the ICC

Just as the p-values of the t-test are based on a t-value, the p-values of ICC-1, -2 and -3 are based on the F-value of the ANOVA models (resp. one-way random, two-way random or two-way mixed ANOVA). The F-value is calculated with the mean variance components described in table 2. If the p-value of the F-test is not significant at the corresponding degrees of freedom, which is based on the number of subjects and raters, there could be insufficient variance between subjects to calculate a reliability coefficient and the coefficient should be looked at with skepticism. Reliability is defined by the amount of agreement between ratings, but is also dependent on the true variance within the sample. True variance in assessment scores can be jeopardized as a consequence of the tendency of

assessors to rate all trainees as average during a live observation. This is also known as the ‘central tendency error’ and has been reported as a problem in in-training evaluation reports (ITER) by some authors.^{23–25} Participants should therefore be stimulated to use the full range of the scales as much as possible and psychosocial barriers for rating trainees as below or above average should be evaluated and managed appropriately.

In ICC-2 and -3, it can additionally be useful to look at the p-value of the F-test for the variance between raters. Opposite to the F-test for the total variance, this p-value should not be significant to indicate there is no significant difference between the assessment scores of raters.

Table 2: Six different ICCs: 3 different models (ICC-1 to ICC-3), all of 2 different types (type 1 and type k). Var = Mean Square (Mean variance). ANOVA = ANalysis Of VAriance. S = Size

ICC	ANOVA	Raters	Subjects	Agreement	Total variance		Error component		S
					ANOVA	ICC	ANOVA	ICC	
1-1	One-way random		Random		Between-groups var	Between- subjects var	Within-group error	Within-subjects error	↓
1-k	One-way random		Random		Between-groups var	Between- subjects var	Within-group error	Within-subjects error	
2-1	Two-way random	Random	Random	Absolute agreement model	Within-subjects var	Between-subjects var	Between-subjects var & Within-subjects error	Between raters var & residual error	
2-k	Two-way random	Random	Random	Absolute agreement model	Within-subjects var	Between-subjects var	Between-subjects var & Within-subjects error	Between raters var & residual error	
3-1	Two-way mixed	Fixed	Random	Consistency agreement model	Within-subjects var	Between-subjects var	Within-subjects error	Error	
3-k	Two-way mixed	Fixed	Random	Consistency agreement model	Within-subjects var	Between-subjects var	Within-subjects error	Error	

2.6 Evaluation of factors influencing the ICC

When two or more assessment methods used by a sample of raters to rate a sample of subjects, seem to differ in terms of reliability, it is sensible to check whether the difference does not originate from a difference in true variance by evaluating the total variance of ratings of the two assessment methods. In figure 2, an example is shown of a hypothetical study in which 3 assessment forms are used: 1) a procedural-based assessment (PBA) for appendectomy, 2) a PBA for hemicolectomy and 3) a global ratings scale (GRS). The scores shown are based on the mean scores of multiple raters. If we assume that the amount of rater error and random error are equal for the appendectomy PBA and the GRS, the reliability of the former would automatically be higher as a consequence of the larger ‘true’ variance in scores (0-95% for the PBA vs. 0-75% for the GRS). For the same reason the PBA for the appendectomy would be more reliable in the earlier stage and the PBA for the hemicolectomy more reliable in the later stage of surgical training.

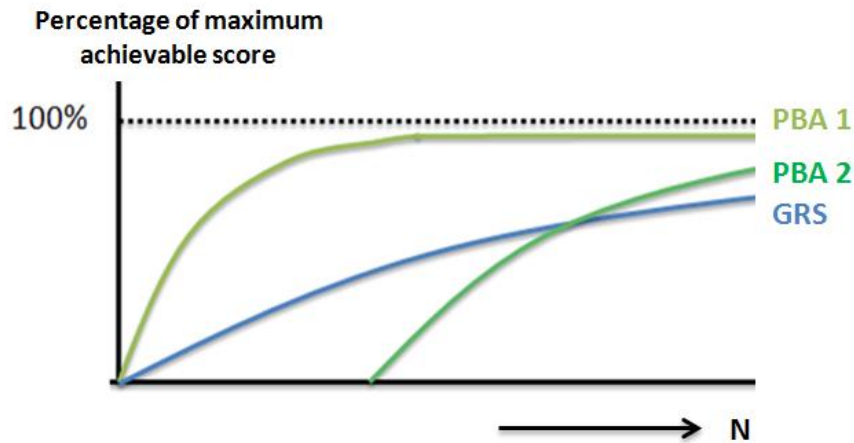


Figure 2: On the x-axis the experience of a trainee and on the y-axis the mean normalized performance level of 2 raters on a procedural-based assessment of an appendectomy (PBA 1), a procedural-based assessment of a hemicolectomy (PBA 2) and a global rating scale (GRS).

Besides true variance, factors like study design, rater background or rater setting can also significantly influence the reliability coefficient in the wrong or right direction. Whether one or more factors had a dominant effect on the outcome can be evaluated with a correlation matrix (Table 3). However, given k raters a correlation matrix consists of $(k-1)*k/2$ ICCs. In the case of a very large quantity of raters it can therefore be valuable to create a geometrical representation by plotting the ICCs between raters as vectors in a graph (Figure 3).²⁶ To achieve a geometrical rendition, the correlations between raters can be calculated into degrees by using the inverse cosine function (\cos^{-1}). The smaller the correlation between variables the larger the angle between the vectors in the graphical representation will be.

Table 3: A correlation matrix of ICC-2,1 of 6 raters performing 20 consecutive assessments. V = Video rater, D = Direct observer. Time point functions as an interaction effect and reduced inter-rater reliability in the group of video raters during the last 5 to 10 assessments (V1-V3).

Raters	Time point			
	1-5	6-10	11-15	16-20
V1,V2,V3	0.90	0.85	0.50	0.12
D1,D2,D3	0.95	0.91	0.89	0.97

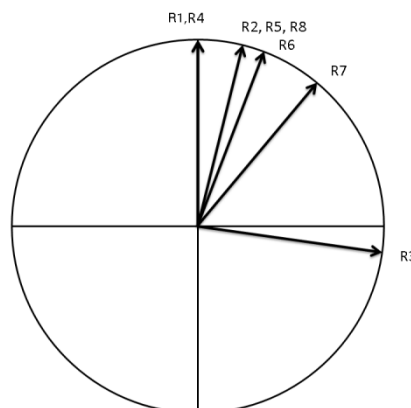


Figure 3: Geographical representation of correlations between 8 raters. R = Rater. The correlation between raters is translated into degrees with the inverse cosine function (\cos^{-1}) to obtain vectors.

3. Study design

There are a number of factors that can result in bias or concerns of applicability of the reliability coefficient reported by studies that address the assessment of surgical skills (Table 4).

Table 4: Issues pertinent to the evaluation of study quality.

Risk of bias	
Study design	Blinding: <ul style="list-style-type: none"> - Subject identity - Handedness - Skin color - Time duration of performance - Voice
	Randomization: <ul style="list-style-type: none"> - Sequence of subject performances - Sequence of assessment forms - Raters
Statistics	If ICC used as measure of inter-rater reliability: chosen ICC model reported?
Concerns of applicability	
Study design	Participants: <ul style="list-style-type: none"> - Rater characteristics - Subject characteristics - Motivation of raters
	Training: <ul style="list-style-type: none"> - Training content described in sufficient detail - Instructions to use full range of scores
Statistics	If ICC used as measure of inter-rater reliability: Appropriate ICC model used

3.1 Transparency

Because there are so many different modalities for calculating reliability, transparency is essential for the assessment of study quality. It can be tempting to use the wrong ICC model, as some models tend to give higher results than others. When Generalizability theory is used, important aspects to report are whether facets are crossed or nested, whether facets are fixed or random, the existence of negative variance components and which software was used. Studies using a Pearson or Spearman correlation coefficient not reporting which of the 6 mathematical different ICCs was used, or not describing the methodological route for estimating generalizability coefficients should be classified as a study of low quality or at risk of bias.^{31,32}

3.2 Randomization and blinding

An assessment method can be evaluated with the raters blinded or not blinded to the identity or experience level of the trainee. A meta-analysis of randomized clinical trials that compared blinded versus non-blinded observer assessments of subjective measurement scales showed that the effect size was exaggerated with 68% in the group of non-blinded studies.³³ Although the number of studies in the surgical literature investigating the effect of concealment of identity is limited, it is reasonable to assume that knowing the identity of the trainee can influence the outcome of assessment. The assessment can theoretically be biased in the positive (Halo effect) or in the negative direction (Devil effect) due to awareness of the identity of the trainee. The risk of introducing bias in the assessment of surgical skills can be avoided in the assessment of surgical skills by using video images restricted to the hands without revealing skin color or handedness, as demonstrated by Vogt et al.³⁴ In the case of laparoscopic surgery, the laparoscope can be used to record a video restricted to images of the inside of the abdomen. Using blinded videos can pose a

problem when items of an assessment incorporate elements of communication of the trainee with the operating team. If verbal elements of communication are to be assessed, a sound recording of the communication can be used to subtitle the video or video parts involving communication can be tagged.

A remaining potential source of bias in the assessment of blinded videos is the time length of the video. Based on the duration of a video, one can estimate roughly the experience level, assuming the difficulty of the task is equal. This can partially be circumvented by using video fragments according to a well-defined protocol. However, editing videos can in turn threaten generalization to the assessment of a whole procedure.

Other potential sources of bias in blinded assessments include the sequence of the performances to be assessed and the sequence of the assessment forms. In drug innovation, there is high level evidence of an exaggeration of the effect size in studies with unclear or inadequate random sequence generation.³⁵ To avoid raters using the sequence of the performances as a source of information for assessment, performances can be randomized. In the case that multiple assessment forms are used simultaneously, raters can develop a raised subconscious or conscious awareness of the strengths and weaknesses during the completion of the first assessment that is transferred to the subsequent assessment. To minimize the chance that the order of the videos or the assessment methods influences reliability, the sequence of assessment can be randomized. An elaborate description of different randomization methods and guidelines on which randomization method to choose according to the study design has been published by Kao et al.³⁶

3.3 Participant characteristics

To avoid concerns of applicability, the included subjects and raters should have characteristics similar to the population of interest. When subjects are chosen, the range of experience levels of subjects should be similar to the range of experience levels in the population in which the measurement instrument is going to be used. For example, demonstrating that an assessment method can reliably assess an expert performance is futile if the assessment method is designed for tracking improvement during training.

A number of authors in surgical education and applied psychology have suggested on the basis of their findings that as raters become more familiar with the assessment method, the reliability of the assessment increases.^{8,38-42} In a recent study in cardiothoracic surgery that investigated the influence of rater training on the reliability of assessment, a dramatic increase in reliability coefficients was observed after training from 0.09-0.48 to 0.80-0.90.²⁰ Therefore, when a new assessment form is tested, or an already validated assessment form is evaluated in another population of raters, training of raters can be necessary to obtain maximum accuracy of assessment scores. Whenever training is relevant for the accuracy of assessment, it is important to report the content of rater training in sufficient detail to allow replication of training in other settings.

Fatigue or a lack of motivation can endanger the accuracy of assessment inside and outside the research context. An assessment form should therefore be able to be completed within a feasible time frame. On the other hand, the internal consistency, measured with Cronbach alpha, is an important aspect of reliability other than inter-rater reliability and tends to increase with the number of items that measure the same trait. The internal consistency rises because measurement errors of the individual items will tend to cancel each other out as the number of items that measure the same trait increases, leading to a more accurate estimate of the measured trait.²⁶ Choosing the total number of items therefore includes finding the optimum balance between feasibility and reliability.

The use of extrinsic rewards can be a valuable instrument to increase interest in participation among surgeons when there is an initial lack of motivation or to optimize commitment and persistence during assessment.⁴³ Care should be taken to avoid diminishing initial intrinsic motivation among volunteering research participants by introducing an (inadequate) external reward, a phenomenon that has been observed in the field of cognitive psychology, and has become known as the 'overjustification effect'.⁴⁴⁻⁴⁶

3.4 Constructivist social-psychological approach

The points raised in this review have primarily been described from a psychometric perspective. Although raters are seen as measurement instruments in the psychometric approach, raters have unique cognitive processes during assessment.⁴⁷ These cognitive processes of assessors are influenced by the acquired knowledge during training within one or more educational institutions, personal operative experiences during their surgical career and the content and characteristics of the interactions with surgical supervisors who trained and supervised them in the skillslab and/or in the OR (socialization). Goovaerts et al. justifiably stated that, although the ratings do not agree from a psychometric standpoint of view, they can all be equally valid and might separately all contribute to a more complete picture of the quality of surgical skills. In our endeavours to objectify surgical skills, we should not forget that the art of surgery can never fully be expressed in something as simple as a number.

4. Conclusions

The public and government have acknowledged that training is a vital aspect of effective patient care and have therefore urged for more objective quality assessment during surgical education. As a consequence, research in surgical education has provided, and will continue to provide, tools to quantify the quality of surgical skills. This review recapitulates on the statistics and study design behind the inter-rater reliability from educational measurement and describes important aspects of the statistics and study design of studies estimating the inter-rater reliability of surgical skills assessment. This paper is aimed to equip surgeon scientists with methods for investigating the inter-rater reliability of subjective assessment and provide designers of surgical training programs and clinical supervisors with the necessary skills for assessing the quality of these studies.

References

1. Hove, P. D. Van, Tuijthof, G. J. M., Verdaasdonk, E. G. G., Stassen, L. P. S. & Dankelman, J. Objective assessment of technical surgical skills. *Br J Surg.* 972–987 (2010).
2. Jelovsek, J. E., Kow, N. & Diwadkar, G. B. Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med. Educ.* 47, 650–673 (2013).
3. Martin, J. A. *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* 84, 273–278 (1997).
4. Niitsu, H. *et al.* Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today.* 43, 271–275 (2012).
5. Hopmans, C. J. *et al.* Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study. *Surgery.* 156, 1078-1088 (2014).
6. Hiemstra, E., Kolkman, W., Wolterbeek, R., Trimbos, B. & Jansen, F. W. Value of an objective assessment tool in the operating room. *Can J Surg.* 54, 116-122 (2011).
7. Vassiliou, M. C. *et al.* A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 190, 107-113 (2005).
8. Kramp, K. H. *et al.* Validity and Reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in Novice Trainees Performing a Laparoscopic Cholecystectomy. *J Surg Educ.* 72, 351-358 (2015).
9. Gallagher, A. G., Ritter, E. M. & Satava, R. M. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc. Other Interv Tech.* 17, 1525–1529 (2003).
10. Karanicolas, P. J. *et al.* Evaluating agreement: conducting a reliability study. *J Bone Joint Surg Am.* 91 Suppl 3, 99–106 (2009).
11. Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull.* 86, 420 (1979).
12. Lahey, M. A., Downey, R. G. & Saal, F. E. Intraclass correlations: There's more there than meets the eye. *Psychol Bull.* 93, 586 (1983).
13. McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1, 30 (1996).
14. Walter, S. D., Eliasziw, M. & Donner, a. Sample size and optimal designs for reliability studies. *Stat Med.* 17, 101–110 (1998).
15. Donner, A. & Eliasziw, M. Sample size requirements for reliability studies. *Stat Med.* 6, 441–448 (1987).
16. Downing, S. M. Reliability: on the reproducibility of assessment data. *Med Educ.* 38, 1006–1012 (2004).
17. Delfino, A. E., Chandratilake, M., Altermatt, F. R. & Echevarria, G. Validation and piloting of direct observation of practical skills tool to assess intubation in the Chilean context. *Med. Teach.* 35, 231-236 (2013).
18. Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D. & Ziv, A. Reliability estimates: behavioural stations and questionnaires in medical school admissions. *Med Educ.* 46, 277–288 (2012).
19. Arain, N. A. *et al.* Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surg Endosc.* 26, 2740–2745 (2012).
20. Lou, X. *et al.* Training less-experienced faculty improves reliability of skills assessment in cardiac surgery. *J Thorac Cardiovasc Surg.* 148, 2491-2496 (2014).
21. Norcini, J. J. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 74, 1088–1090 (1999).
22. Harvill, L. M. Standard Error of Measurement. *Educ Meas Issues Pract.* 33–41 (1991).

23. Feldman, L. S., Hagarty, S. E., Ghitulescu, G., Stanbridge, D. & Fried, G. M. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *J Am Coll Surg.* 198, 105–110 (2004).
24. Mclaughlin, K., Vitale, G., Coderre, S., Violato, C. & Wright, B. Clerkship evaluation—what are we measuring? *Med Teach.* 31, e36–e39 (2009).
25. Gray, J. D. Global rating scales in residency education. *Acad Med.* 71, S55–S63 (1996).
26. Cooper, C. *Individual Differences.* (Routledge, 2002).
27. Winkel, C. P., Reznick, R. K., Cohen, R. & Taylor, B. Reliability and construct validity of a Structured Technical Skills Assessment Form. *Am J Surg.* 167, 423–427 (1994).
28. Scott, D. J. *et al.* Measuring Operative Performance after Laparoscopic Skills Training: Edited Videotape versus Direct Observation. *J. Laparoendosc. Adv Surg Tech.* 10, 183–190 (2000).
29. Sidhu, R. S., Vikis, E., Cheifetz, R. & Phang, T. Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? *Am J Surg.* 191, 677–681 (2006).
30. Melchioris, J. *et al.* Preparing for Emergency: A Valid, Reliable Assessment Tool for Emergency Cricothyroidotomy Skills. *Otolaryngol. -- Head Neck Surg.* 152, 260–265 (2014).
31. Krebs, David E. Opinions and Comments. *Rehabilitation* 67, 22314–22314 (1987).
32. Kottner, J. *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol.* 64, 9096–9106 (2010).
33. Hróbjartsson, A. *et al.* Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *Can Med Assoc J.* 185, E201 (2013).
34. Vogt, V. Y., Givens, V. M., Keathley, C. A., Lipscomb, G. H. & Summitt, R. L. Is a resident's score on a videotaped objective structured assessment of technical skills affected by revealing the resident's identity? *Am J Obstet Gynecol.* 189, 688–691 (2003).
35. Savović, J. *et al.* Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: Combined analysis of meta-epidemiological studies. *Health Technol Assess. (Rockv).* 16, 1–81 (2012).
36. Kao, L. S., Tyson, J. E., Blakely, M. L. & Lally, K. P. Clinical Research Methodology I: Introduction to Randomized Trials. *J Am Coll Surg.* 206, 361–369 (2008).
37. Korndorffer, J. R., Kasten, S. J. & Downing, S. M. A call for the utilization of consensus standards in the surgical education literature. *Am J Surg.* 199, 99–104 (2010).
38. Vassiliou, M. C. *et al.* Evaluating Intraoperative Laparoscopic Skill: Direct Observation Versus Blinded Videotaped Performances. *Surg Innov.* 14, 211–216 (2007).
39. Schijven, M. P. *et al.* Transatlantic comparison of the competence of surgeons at the start of their professional career. *Br J Surg.* 97, 443–449 (2010).
40. Dath, D. *et al.* Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc.* 18, 1800–1804 (2004).
41. Lievens, F. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *J Appl Psychol.* 86, (2001).
42. Matsuda, T. *et al.* Reliability of Laparoscopic Skills Assessment on Video: 8-Year Results of the Endoscopic Surgical Skill Qualification System in Japan. *J Endourol.* 28, 1374–1378 (2014).
43. Gneezy, U. & Rustichini, A. Pay Enough or Don't Pay at All. *Q J Econ.* 115, 791–810 (2000).
44. Cameron, J., Banko, K. M. & Pierce, W. D. Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *Behav Anal.* 24, 1–44 (2001).
45. Frey, B. & Goette, L. Does pay motivate volunteers? 22 (1999). doi:10.3929/ethz-a-004372692
46. Mellström, C. & Johannesson, M. Crowding out in blood donation: Was Titmuss right? *J Eur Econ Assoc.* 6, 845–863 (2008).
47. Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T. & Muijtjens, A. M. M. Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Adv Heal Sci Educ.* 12, 239–260 (2007).

Chapter 7

**VALIDITY AND RELIABILITY OF GLOBAL
OPERATIVE ASSESSMENT OF LAPAROSCOPIC
SKILLS (GOALS) IN NOVICE TRAINEES
PERFORMING A LAPAROSCOPIC
CHOLECYSTECTOMY**

K.H. Kramp, M.J. Det, C. Hoff, B. Lamme, N.J.G.M. Veeger, Jean-Pierre E.N. Pierie

Journal of Surgical Education 2015;72(2):351-358

Abstract

Purpose: Global Operative Assessment of Laparoscopic Skills (GOALS) assessment has been designed to evaluate skills in laparoscopic surgery. A longitudinal blinded study of randomized video fragments was conducted to estimate the validity and reliability of GOALS in novice trainees.

Methods: Ten trainees each performed 6 consecutive laparoscopic cholecystectomies. Sixty procedures were recorded on video. Video fragments of 1) opening of the peritoneum, 2) dissection of Calot's Triangle and achievement of Critical View of Safety (CVS) and 3) dissection of the gallbladder from the liver bed were blinded, randomized and rated by two consultant surgeons with GOALS. Also, a grade was given for overall competence. The correlation of GOALS with live observation OSATS scores (Objective Structured Assessment of Technical Skills) was calculated. Construct validity was estimated with the Friedman two-way analysis of variance by ranks and Wilcoxon signed-rank test. The inter-rater reliability was calculated with the absolute and consistency agreement two-way random effects model intra-class correlation coefficient.

Results: A high correlation was found between mean GOALS score ($r = 0.879$, $p = 0.021$) and mean OSATS score. The GOALS score increased significantly across the 6 procedures ($p = 0.002$). The trainees performed significantly better on their sixth when compared with their first cholecystectomy ($p = 0.004$). Consistency agreement inter-rater reliability was 0.37 for the mean GOALS score ($p = 0.002$) and 0.55 for overall competence ($p = 0.002$) of the 3 video fragments.

Conclusion: The validity observed in this randomized blinded longitudinal study supports the existing evidence that GOALS is a valid tool for assessment of novice trainees. Compared to other studies a low reliability was found in this study.

Introduction

Objective assessment of technical skills of surgical trainees is an important topic in the field of surgical education. To provide a valid and reliable tool in the assessment of surgical skills, Martin et al. developed a global rating scale in the late 1990s¹, currently known as the OSATS (Objective Structured Assessment of Technical Skills). OSATS has been implemented in many academic centers to measure operative performances in the operating theater and provide feedback to the trainee. Although the OSATS is considered to be a validated tool for global assessment of operative competence^{2,3}, there was no equivalent for laparoscopic surgery. Since laparoscopic surgery is the standard for an increasing list of procedures, there was a need for a valid and reliable assessment tool that addresses the specific requirements of laparoscopic surgery. Laparoscopic surgery involves a man-machine environment that requires the ability to work with a 2-dimensional view, decreased degrees of freedom and reduced tactile feedback. Furthermore, the surgeon is challenged by the fulcrum effect; inversion and scaling of movements of the parts of the instruments inside the abdomen. To evaluate these skills Vassiliou et al. developed GOALS (Global Operative Assessment of Laparoscopic Skills), a non-procedure specific assessment tool that can be applied to any procedure in minimally invasive surgery (MIS).⁴

Rasmussens' model of human behavior can be used to describe different levels that have to be achieved in laparoscopic skill training to obtain competency in MIS.⁵ In the first level the trainee acquires skill-based behavior by learning automated sensory-motor patterns. It has been shown that these skills can be improved on a virtual reality simulator.⁶ In the early post-simulator development phase, learned sensory-motor patterns are calibrated to the MIS environment while rule- and knowledge-based behaviors are acquired. Moore & Bennet demonstrated that the risk of complications is approximately 1.7% in the first laparoscopic cholecystectomy and decreases to 0.7% after 5 cases.⁷ Although much has changed in the education of trainees since then, this novice development stage can still be considered as one of the most important learning phases in guiding surgical trainees to competency in performing a laparoscopic cholecystectomy. This study was conducted to explore the validity and reliability of using GOALS for video-assessment of laparoscopic cholecystectomy in this critical learning phase.

Method

Participants and patient selection

Ten surgical residents in their first (N=4) and second (N=6) year of training were recruited for a training curriculum in laparoscopic cholecystectomy. Only trainees who had attended less than 5 laparoscopic procedures and had no experience with performing a laparoscopic cholecystectomy were included. A minimum of 6 months experience with open surgery was a prerequisite to participate in the study. After a basic laparoscopic skills training the trainees performed 6 laparoscopic cholecystectomies in the OR under the supervision of one of the three participating surgeons. All patients included in the study had uncomplicated symptomatic gallstone disease. All patients gave informed consent before undergoing surgery.

Basic laparoscopic skills training

Basic laparoscopic skills were acquired on the SIMENDO laparoscopy trainer (Simendo, Rotterdam, the Netherlands). The intention of the SIMENDO simulator training is to teach trainees a specified level of basic automated sensory-motor patterns required for safe participation in laparoscopic procedures in humans.

Direct observation: OSATS assessment

The OSATS was developed by Martin et al. in 1997 and is currently the standard method for the assessment of surgical skills.¹ The OSATS consists of 7 items: 1) respect for tissue, 2) time and motion, 3) instrument handling, 4) knowledge of instruments, 5) use of assistants, 6) flow of operation and 7) knowledge of the procedure. Each item was scored as generally used in the Dutch surgical training program on a 10-point scale.

The three supervising surgeons that randomly supervised the operations used the OSATS to assess the laparoscopic performance of the trainees. Because OSATS assessment is an integral part of the surgical curriculum in the Netherlands, the surgeons had used the OSATS frequently in the past to assess trainees. The surgeons were uninformed about the number of procedures the trainee performed previously, but not blinded to the identity of the trainee.

To determine whether the increase in OSATS is mainly caused by non sensory-motor skill acquisition, the OSATS-sm (OSATS-sensory-motor) was calculated by summing the items 1, 2, 3 and 6 of the OSATS form.

Video assessment: GOALS and overall competence

The GOALS assessment form contains 6 items. Four items represent domains of technical competence in laparoscopic surgery: 1) depth perception, 2) bimanual dexterity, 3) efficiency and 4) tissue handling. The 5th item is used to rate the autonomy of the subject. Only parts of the video in which the trainee performed as operating surgeon were edited so the item autonomy was therefore left out of the GOALS form. The 6th item, level of difficulty, was added by Chang et al. to also take into account any difference in difficulty of the procedure.⁸

To be able to compare GOALS with the modified 10-point version of the OSATS global rating scale used in our institution, the items on the GOALS form were converted to a 10-point scale. Complementary to the GOALS items, a grade for overall competence was rated on a 10-point scale for each video fragment. It has been shown that transformation of a 5-point scale to a 10-point scale does not significantly influence the data characteristics besides a slightly decrease in the scores with respect to the maximum achievable score.⁹

During every procedure a video was recorded with the laparoscopic camera and audio was recorded with 2 microphones; one attached to the trainee and one to the supervising surgeon. The videos were divided into 3 sections: 1) opening of peritoneum, 2) dissection of Calot's Triangle and achievement of CVS and 3) dissection of the gallbladder from the liver bed. The audio material was used to identify the sections in which the trainee was acting as the operating surgeon. When a

supervising surgeon took over the procedure, that part was cut from the video. The video fragments were terminated after 5 minutes or when a section was completed. Subsequently, the videos were muted so the raters were blinded for the performing trainee and the supervising surgeon. After editing and removal of the audio, the order for video assessments was randomly set on the basis of participating trainee and number of cholecystectomies performed while the order of the video fragments was maintained. Each individual video fragment was rated by two consultant surgeons who were involved in the training program for laparoscopic surgery (Figure 1).

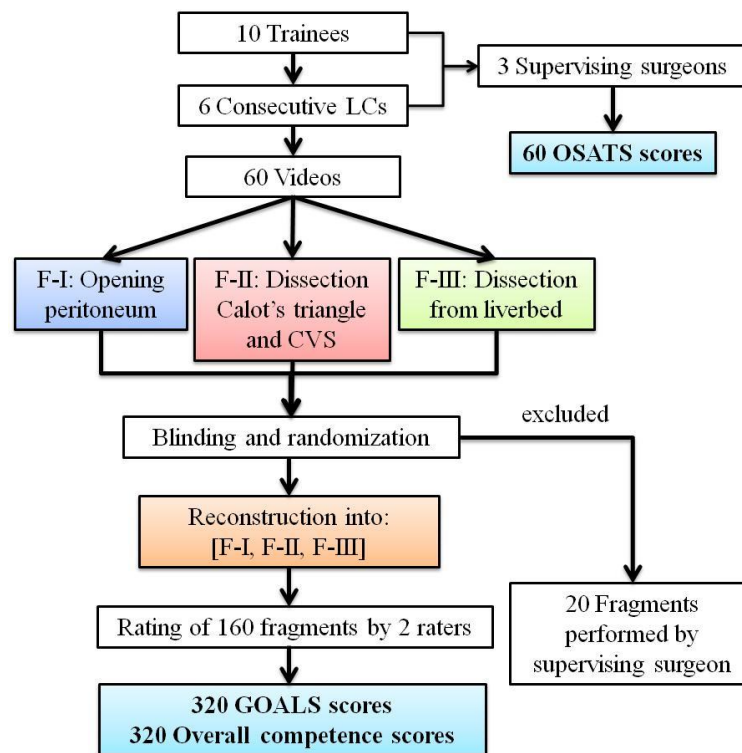


Figure 1: Workflow. F=Video fragment; LC=laparoscopic cholecystectomy

Statistical analysis

The usefulness of a measurement tool is dependent on the degree that it measures what it is supposed to measure (validity) and the accuracy of those measurements (reliability). The GOALS scores were used to calculate construct validity (increase in performance score with increase in caseload), concurrent validity (correlation with the OSATS) and interrater reliability (absolute and consistency agreement between two raters). SPSS 20.0.0.1 (SPSS, Chicago, IL, USA) was used in all analysis. Statistical significance was defined as $p < 0.05$.

Validity

To estimate concurrent validity, the correlation between mean GOALS score and OSATS score of the supervising surgeon was calculated with the Pearson's r correlation coefficient. The Friedman two-way analysis of variance by ranks was used to estimate the construct validity. In addition, the performance on the first was compared with the performance on the sixth cholecystectomy with the Wilcoxon signed-rank test.

Reliability

The intra-class correlation coefficient (ICC) was used to calculate the reliability. Because the ability to estimate progression is the most important aspect of the learning trajectory of the trainees in our study sample, we were interested in the commonly used absolute agreement, but also in the

consistency agreement inter-rater reliability between the two raters. Therefore, the absolute agreement two-way random effects model for single measures (AA-ICC 2,1) and the consistency agreement two-way mixed effects model for single measures (CA-ICC 3,1) of the intra-class correlation coefficient were chosen.¹⁰⁻¹²

The mean total GOALS score, the mean items score and mean overall competence score of 3 video fragments was compared between the two raters. Values used for ordinal classification of the inter-rater reliability are always arbitrary in nature and should be adjusted to the purpose of the measurement instrument. Because GOALS would primarily be used for formative assessment in our study population and not for high stakes examination, cut-off points for the ICC were chosen as 'moderate' (0.21 to 0.40), 'reasonable' (0.41 to 0.60), 'good' (0.61 to 0.80) and 'almost perfect' (0.81 to 1.00).^{13,14}

Results

Measurements

Sixty laparoscopic cholecystectomies were successfully recorded. A total of 160 video fragments were blinded, randomized and rated by 2 raters. Twenty video fragments could not be rated due to intervention of the supervising surgeon. There were no technical problems. The yield was 320 measurements (Figure 1).

As presented in table 1, the mean OSATS score of the two raters was 20.2±8.5 at procedure 1 and increased to 43.5±6.6 at procedure 6. The mean OSATS-sm increased from 10.5±4.1 at procedure 1 to 23.6±4.2 at procedure 6. The mean GOALS score of the two raters increased from 20.0±4.8 at procedure 1 to 23.7±4.3 at procedure 6. The mean overall competence score of the two raters was 4.6±1.1 at procedure 1 and 5.4±1.0 at procedure 6.

Table 1: Mean OSATS score and mean OSATS-sm score (item 1, 2, 3 and 6 from OSATS) per caseload.

Procedure	1	2	3	4	5	6	p	p(Δ1-6)
OSATS-sm	10.5±4.1	14.4±3.5	17.8±5.7	18.2±5.7	19.7±4.8	23.6±4.2		
OSATS	20.2±8.5	27.5±7.3	34.2±10.0	34.9±11.3	37.6±6.0	43.5±6.6	<0.001*	0.008*

* Statistical significant

Validity

A high correlation between mean GOALS score and mean OSATS score was observed ($r=0.879$, $p = 0.021$).

The OSATS scores increased significantly with caseload ($p < 0.001$) and there was a significant difference between the OSATS scores of the trainees measured in the first versus sixth operation ($p = 0.008$) (Figure 2). Approximately 50% of the total increase in OSATS scores consisted of sensory-motor items (Table 1).

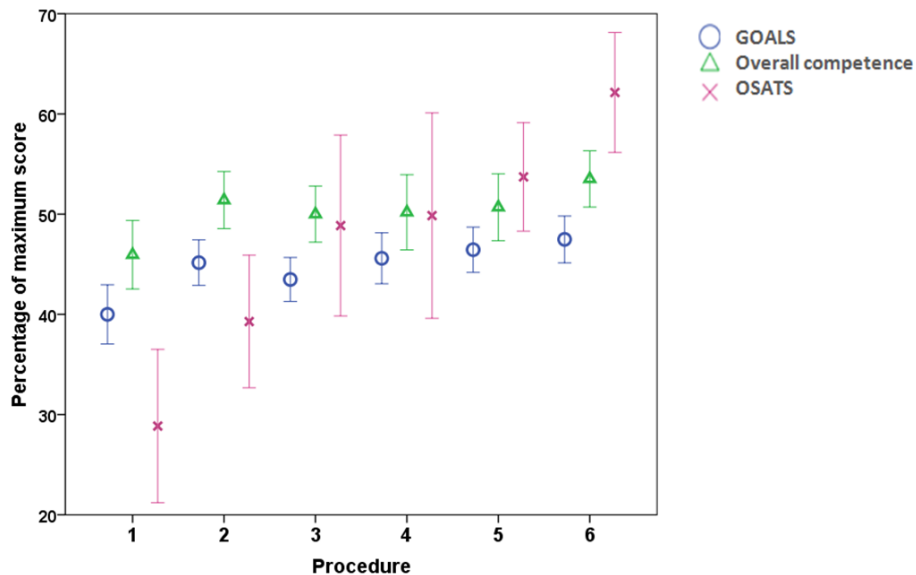


Figure 2: Increase in OSATS score, GOALS score and overall competencescore. The difference in OSATS score, GOALS score and overall competence score in the 6 consecutive procedures was significant ($p = 0.008$, $p = 0.002$ and $p = 0.016$). Error bars indicate 95% confidence intervals.

The GOALS scores increased significantly with caseload ($p < 0.001$) and there was a significant difference between the GOALS scores of the trainees measured in the first versus sixth operation ($p =$

0.004) (Table 2). The overall competence also increased significantly with caseload ($p = 0.016$) and between the first and sixth operation ($p = 0.003$) (Table 2).

The GOALS scores and overall competence scores of the video fragments only showed a significant increase in the video fragment of the dissection of Calot's Triangle (Table 2).

Table 2: Number of fragments, mean GOALS score per fragment, mean GOALS score for 3 fragments and mean overall competence score per caseload.

Number of fragments								
Procedure	1	2	3	4	5	6		
N total=160	21	29	27	27	29	27		
GOALS score								
Procedure	1	2	3	4	5	6	p	p(Δ 1-6)
F1: Opening of the peritoneum	18.9±5.1	20.9±3.9	20.3±4.6	19.3±3.9	24.6±5.6	23.7±5.1	0.063	0.208
F2: Dissection of Calot's Triangle and achievement of CVS	19.4±4.9	22.7±4.1	22.3±3.8	25.3±4.3	22.5±3.9	23.9±4.2	0.005*	0.011*
F3: Dissection from the liver bed	21.1±4.6	24.0±4.4	22.6±3.4	24.6±3.4	22.7±3.0	23.6±3.7	0.129	0.447
Overall competence								
Procedure	1	2	3	4	5	6	p	p(Δ 1-6)
F1: Opening of the peritoneum	4.5±1.2	4.7±1.2	4.7±1.2	4.1±1.3	5.2±1.6	5.2±1.2	0.525	0.305
F2: Dissection of Calot's Triangle and achievement of CVS	4.4±1.2	5.3±1.0	5.0±1.0	5.4±1.3	4.9±1.1	5.3±1.0	0.021*	0.030*
F3: Dissection from the liver bed	4.8±1.0	5.4±0.9	5.3±0.9	5.7±1.0	5.1±1.1	5.7±0.7	0.113	0.068

P-values are based upon the Friedman two-way analysis of variance by ranks and the Wilcoxon signed-rank test.* Statistical significant

Reliability

Table 3 shows the AA-ICC and CA-ICC of the mean total GOALS score, the mean GOALS items score and mean overall competence of the 3 video fragments. The AA-ICC and CA-ICC of the mean GOALS score were moderate (0.37; $p = 0.002$, 0.37; $p = 0.002$) (Figure 3). The highest AA-ICC was found for the item 'efficiency' (0.47; $p < 0.001$) and the lowest for the item 'level of difficulty' (0.22; $p < 0.001$). The highest CA-ICC was found for the item 'level of difficulty' (0.55, $p < 0.001$) and lowest for the item 'bimanual dexterity' (0.27; $p = 0.019$).

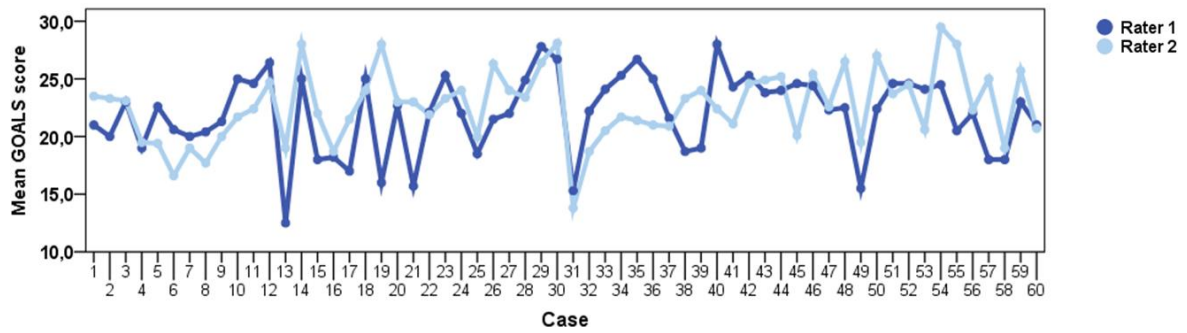


Figure 3: Inter-rater reliability of mean GOALS score of fragment 1 to 3 between rater 1 and 2.

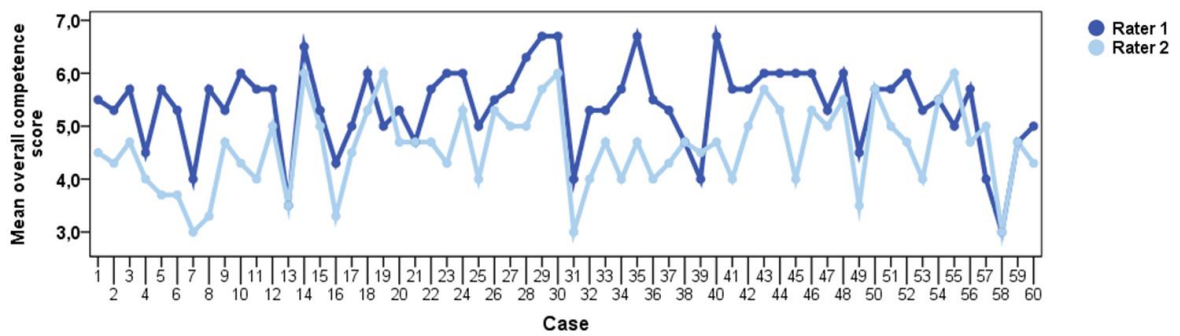


Figure 4: Inter-rater reliability of mean overall competence score of fragment 1 to 3 between rater 1 and 2.

Table 3: The AA- and CA-ICC of the mean GOALS score of the items, the mean total GOALS score and overall competence score of the assessment of three video fragments of a procedure of one trainee (N=320).

Item	Domain	AA-ICC (2,1)	CA-ICC (3,1)
1	Depth perception	0.23	0.33
2	Bimanual dexterity	0.26	0.27
3	Efficiency	0.47	0.47
4	Tissue handling	0.34	0.33
5	Level of difficulty	0.22	0.55
Overall competence score		0.36	0.55
GOALS score		0.37	0.37

All ICC-values were statistical significant ($p < 0.05$)

Discussion

Objective assessment of surgical trainees is necessary to ensure professional standards are being met in the operating room. In contrast to OSATS, GOALS contains specific criteria for minimally invasive surgery. In this longitudinal study GOALS was used to assess blinded randomized video fragments of a laparoscopic cholecystectomy.

Six consecutive cholecystectomies performed by 10 trainees were recorded. Out of the video recordings 3 fragments were edited to produce a total of 160 video fragments which were assessed by two blinded raters previously unexposed to GOALS. The significant increase in mean GOALS score across the 6 procedures and the high correlation with the mean OSATS score indicate that GOALS is a valid instrument for assessment of laparoscopic surgical skills.

Validity

Earlier studies have shown that GOALS can distinguish surgeons of varying skill level^{4,8,16,17}, but there are only two blinded studies that evaluated GOALS.^{8,17} The first study did not use repeated measurements of the same trainees, leaving room for individual differences between trainees to influence the results.¹⁷ The second study was a blinded study that used 2 videos: one of a novice and one of an expert. Although the study was blinded, the high difference in skill performance of the two videos was derivable from the video duration time (55 min. vs. 15 min.).⁸ Both studies indicate that assessors can thus distinguish a novice from an expert with GOALS, but provide no longitudinal information about the learning curve measured with GOALS. In our study the increase in performance was tracked with repeated measurements of an identical group of trainees with no prior in vivo laparoscopic experience to highlight the value of GOALS assessment and its implementation in surgical training programs. Furthermore, the video fragments were not only blinded, but also randomized. Raters were therefore not only unaware of the identity of the trainee, but also of the number of cholecystectomies performed previously.

Although the results indicate statistical significant construct validity, the difference in mean GOALS score between the first and sixth procedure was minimal (7%). This may be caused by several factors. First, the high score of approximately 40% of the maximum score in the first procedure suggests that the raters should have been encouraged more to use the full range of the items on the GOALS form. Second, it could be caused by a 'real' high level of sensory-motor skill level in the first procedure achieved through simulator training in the basic laparoscopic training course, although the mean percentage of the maximum score in the first procedure of the OSATS (29%) and OSATS-sm (26%) do not support this. A third possible cause is the absence of feedback to the trainee based on the GOALS items, as was done with the OSATS. Feedback gives the trainee the opportunity to strengthen his or her weaknesses and achieve a higher score in the assessment of a subsequent performance.

In this study a significant increase in mean GOALS score and mean overall competence score was only observed in the video fragment of the dissection of Calot's Triangle and achievement of CVS. These results are consistent with those observed by Aggarwal et al. with motion tracking data.¹⁵ Aggarwal et al. found a statistical significant difference in time taken, total path length and number of movements in the video fragment of the dissection of Calot's Triangle between a novice and experienced group. They did not find any difference with motion tracking data in clipping and cutting of the cystic artery, clipping and cutting of the cystic duct and in the dissection of the gallbladder from the liver bed in path length and number of movements. The most likely explanation for these observations is that the dissection of the Calot's Triangle is the hardest step to complete. As a result, it is probably the most sensitive procedural step for operative performance measurements such as GOALS assessment, overall competence scores or motion tracking data.

Reliability

A low reliability was observed in the mean of the three video fragments of one procedure performed by a trainee (0.37). The low ICC means that a low percentage (37%) of the difference between ratings is attributable to true variance and the remaining variance is attributable to other sources.

There are multiple factors that can influence the reliability in assessments. An important factor is the training of the raters in the assessment method. The lowest reliability of GOALS was reported by Vassiliou et al.¹⁶ In this study Vassiliou et al. compared direct observation ratings with blinded videotape ratings. They found an ICC of 0.39 when the scores of one of the video raters were compared with those of 2 direct raters. They ascribe the ICC of 0.39 to the video raters' lack of previous exposure to GOALS. Vassiliou et al. also describe a video rater that was in like manner unexposed to the GOALS, but attained an ICC of 0.76 when his scores were compared with the 2 direct observations. This video rater reported to have invested a considerable amount of time in getting comfortable with the assessment method by watching all the videos beforehand and watching videos multiple times during the assessment.¹⁶ According to the authors, these findings suggest that training in GOALS assessment might be necessary before reliable GOALS scores can be achieved. Matsuda et al. had similar findings in their study of the Endoscopic Surgical Skill Qualification System in Japan; the amount of exposure to their assessment method correlated significantly with the reliability of the ratings. They stated that long-term experience with their assessment method is necessary to perform reliable skill assessments.²² The results of this study might indicate likewise that the interrater reliability is jeopardized when GOALS is used without proper instructions and/or training of the raters.

A second contributing factor lies in the calculation used for estimating the ICC. The ICC harbors the variance within the sample to calculate the reliability. As the estimated true variance on the basis of between-subject variance decreases the calculated ICC automatically tends to decrease.¹⁷

A third explanation could be in the scale used in the GOALS form. Some authors state that attaining an absolute agreement reliability of 0.80 is one of the major inherent difficulties of using a Likertscale.¹⁴

Finally, although raters involved in surgical education are probably inclined to invest energy and time in the assessment, their motivation may be threatened by mental fatigue or time pressure and therefore lead to unreliable measurements. Practical consequences of this may be that video assessments are limited to a particular section of the operation or raters are rewarded to guarantee sufficient motivation.

Limitations of this study

Although our measurements indicate that GOALS has significant construct and concurrent validity when assessing novice trainees, some limitations should be kept in mind. First, different methods were used for OSATS and GOALS assessment; OSATS assessment was performed with direct observation and GOALS assessment with video fragments. Second, we used a total of 320 GOALS assessments to measure the improvement in surgical skills. Our large sample size probably disguised the low inter-rater reliability and made it possible to establish validity. Therefore, the question remains whether the validity also exists in the operating theater when the measurement of skill level is based a smaller sample of measurements. On the other hand, the validity could be higher because the item autonomy is included and/or the rater assesses the whole procedure instead of only fragments. Third, assessing a consecutive series of 6 identical procedures probably seldom takes place during a residency. In most cases there is an interval of learning without formal assessment of the trainee. Fourth, although the raters were consultant surgeons that were familiar with assessing trainees, we did not identify whether there existed a difference in the perception of what can be defined as 'good' or 'bad' laparoscopic skills.

In the field of minimally invasive surgery, there is a demand for objective assessment of professional skills in order to meet increasing political and public demands. The availability of an objective

assessment method gives educators the opportunity to certify trainees according to their abilities. Certification enables a formal, transparent and objective identification of trainees who are able to complete laparoscopic procedures independently, skillfully and most important, safely. The OSATS can be considered as an option, but the results of recent studies have raised concerns about the objectivity of the OSATS and some authors therefore reject the idea that OSATS can function as an instrument for summative assessment.^{20,21} GOALS could be a better alternative to the OSATS for this purpose. Therefore, it is important to mention that the reliability found in our study sample cannot be generalized to trainees in higher ranges of surgical skill level.

Conclusions

In conclusion, this randomized blinded longitudinal study supports the existing evidence that GOALS has construct and concurrent validity for assessment of novice trainees performing a laparoscopic cholecystectomy. The reliability observed in this study was low compared to the reliability found in other studies.

References

- 1 Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278.
- 2 Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today*. 2013;43(3):271-275.
- 3 Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. *Can J Surg*. 2011;54(2):116-122.
- 4 Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113.
- 5 Wentink M, Stassen LP, Alwayn I, Hosman RJ, Stassen HG. Rasmussen's model of human behavior in laparoscopy training. *Surg Endosc*. 2003;17(8):1241-1246.
- 6 Ikonen TS, Antikainen T, Silvennoinen M, Isojärvi J, Mäkinen E, Scheinin TM. Virtual reality simulator training of laparoscopic cholecystectomies - a systematic review. *Scand J Surg*. 2012;101(1):5-12.
- 7 Moore MJ, Bennett CL. The learning curve for laparoscopic cholecystectomy. The Southern Surgeons Club. *Am J Surg*. 1995;170(1):55-59.
- 8 Chang L, Hogle NJ, Moore BB, et al. Reliable assessment of laparoscopic performance in the operating room using videotape analysis. *Surg Innov*. 2007;14(2):122-126.
- 9 Dawes J. Do Data Characteristics Change According to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Int J Market Res*. 2008;50(1): 61-77.
- 10 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psych Meth*. 1996;1(1):30-46.
- 11 Shrout PE, Fleis JL. Intraclass correlation: Uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
- 12 Krebs DE. Declare your ICC type. *Phys Ther*. 1986;66:1431.
- 13 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
- 14 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.
- 15 Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. 2007;245(6):992-999.
- 16 Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204(2):308-313.
- 17 Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov*. 2007;14(3):211-216.
- 18 Watkins MP. Foundations of clinical research – Application to practice, 3th edition. Portney LG, Pearson Education, Inc., New Jersey 07458.
- 19 Hiemstra E, Khaledian NH, Trimbos JBMZ, Jansen FW. Implementation of OSATS in the residency program: a benchmark study. *Submitted*.
- 20 Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg*. 2010;97(7):972-987
- 21 Richard R, Deci EL. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Cont Ed Psych*. 2000;25(1):54-67.
- 22 Matsuda T, Kanayama H, Ono Y, et al. Reliability of Laparoscopic Skills Assessment on Video: 8-Year Results of the Endoscopic Surgical Skill Qualification System in Japan. *J Endourol*. 2014 Jul 16. [Epub ahead of print]

Chapter 8

**VALIDITY, RELIABILITY AND SUPPORT FOR
IMPLEMENTATION OF INDEPENDENCE-SCALED
PROCEDURAL ASSESSMENT IN LAPAROSCOPIC
SURGERY**

Kelvin H. Kramp, Marc J. van Det, Nic J.G.M. Veeger, Jean-Pierre E.N. Pierie

Surgical Endoscopy 2016;30(6):2288-2300

Abstract

Background: There is no widely used method to evaluate procedure specific laparoscopic skills. The first aim of this study was to develop a procedure based assessment method. The second aim was to compare its validity, reliability and feasibility with currently available global rating skills (GRSs).

Method: An independence-scaled procedural assessment was created by linking the procedural key steps of the laparoscopic cholecystectomy to an independence scale. Subtitled and blinded videos of a novice, an intermediate and an almost competent trainee were evaluated with GRSs (OSATS and GOALS) and the independence-scaled procedural assessment by 7 surgeons, 3 senior trainees and 6 scrub nurses. Participants received a short introduction to the GRSs and independence-scaled procedural assessment before assessment. The validity was estimated with the Friedman and Wilcoxon test and the reliability with the intra-class correlation coefficient (ICC). A questionnaire was used to evaluate user opinion.

Results: Independence-scaled procedural assessment and GRS scores improved significantly with surgical experience (OSATS $p = 0.001$, GOALS $p < 0.001$, independence-scaled procedural assessment $p < 0.001$). The ICCs of the OSATS, GOALS and independence-scaled procedural assessment were resp. 0.78, 0.74 and 0.84 among surgeons. The ICCs increased when the ratings of scrub nurses were added to those of the surgeons. The independence-scaled procedural assessment was not considered more of an administrative burden than the GRSs ($p = 0.692$).

Discussion/Conclusion: A procedural assessment created by combining procedural key steps to an independence scale is a valid, reliable and acceptable assessment instrument in surgery. In contrast to the GRSs, the reliability of the independence-scaled procedural assessment exceeded the threshold of 0.8, indicating that it can also be used for summative assessment. It furthermore seems that scrub nurses can assess the operative competence of surgical trainees.

Introduction

Traditionally, assessment of trainees is based on objective but unreliable measures of surgical skills such as blood loss, operation time and perioperative complications. As an alternative Martin et al. developed the Objective Surgical Assessment of Technical Skills (OSATS).¹ The OSATS has been validated in a series of studies and has become the golden standard for structured feedback towards trainees.²⁻⁵ However, in the last decennia laparoscopic surgery has become the standard of care for an increasing list of procedures. In contrast to open surgery, performing laparoscopic surgery requires the ability to work with a 2-dimensional view, decreased degrees of freedom, reduced tactile feedback and the fulcrum effect (inversion and scaling of movements of the parts of the instruments inside the abdomen). Therefore, Vassiliou et al. developed Global Operative Assessment of Laparoscopic Skills (GOALS), a non-procedure specific assessment tool that can be used to assess procedures in minimal invasive surgery (MIS).^{6,7} Although global rating scales (GRSs), such as the OSATS and GOALS, are useful tools for formative assessment (feedback during learning in low-stakes evaluation), a systematic review conducted by Van Hove et al. demonstrated a lack of high level evidence that these and other GRSs are reliable enough for summative assessment (assessment of learning in high-stakes examinations) in the OR.⁴ Furthermore, a survey among gynecological residents and gynecologists indicated that the OSATS was not considered an objective instrument for assessment.⁵ In another survey, conducted by Beard et al. among clinical supervisors and trainees, the greatest number of negative responses was related to the use of OSATS for summative assessment.⁸ The insufficient reliability and the negative responses about the objectivity of the OSATS in surveys are shortcomings that have been used as arguments to prohibit the use of the GRSs as tools for summative assessment in surgical education.^{4,5,8}

Procedural assessment has been proposed as an alternative to GRSs.⁸ A procedural assessment method could enable clinicians to provide procedural specific feedback and, in contrast to the GRSs, could facilitate examination in the performance of a procedure. In order to be useful for these purposes it should comply with three requirements. First, it should be a valid measure of improvement in performance level in a procedure. Second, to facilitate summative assessment, it should be a highly reliable tool in identifying trainees who can safely perform uncomplicated procedures without supervision. Third, it should have enough support from trainees and supervising surgeons to make implementation into clinical practice feasible. To our knowledge, there is no widely used procedural assessment yet that meets all these demands. Hence, we had three research goals:

- 1)** To create a procedural assessment for a procedure that is routinely performed with minimal invasive surgery, the laparoscopic cholecystectomy (LC).
- 2)** To estimate the validity, reliability and support for implementation of this assessment method.
- 3)** To compare the validity, reliability and support for implementation of the procedural assessment with that of the already existing GRSs.

Materials and methods

Development of the independence-scaled procedural assessment

A procedural assessment for the LC was developed in two phases. The first phase has recently been published and consists of twenty-one experts from the North-East Surgical School of The Netherlands that participated in an anonymous survey about the procedural key steps of the LC.⁹

In the second phase, conducted in the present study, the key procedural steps were linked to a rating scale published by Glarner et al. to create an independence-scaled procedural assessment for the LC.¹⁰ This rating scale was chosen because it was observed that in the learning situation supervising surgeons aimed to find a balance between creating the optimal learning experience for the trainee and guarding the patient-safety and flow throughout the operation. They attempted to achieve this goal with: 1) verbal guidance and 2) takeovers. Verbal guidance, consisting of instructions and corrections, were given to optimize surgical behavior. If verbal guidance insufficiently corrected the behavior of the trainee, supervising surgeons tend to take over one or both instruments to guard the safety and flow of the procedure. The independence based assessment model used by Glarner et al. connects to this balance between patient-first mentality and creating the optimal learning environment. It is different from a Likert-type scale in that the frequency of verbal guidance and takeovers are used to quantify the quality of surgical skills.

The independence-scaled procedural assessment for the LC was used in a pilot experiment in the OR and iteratively adjusted on the basis of feedback from trainees and supervising surgeons. The final version of the independence-scaled procedural assessment is displayed in Figure 1.

Procedural assessment Laparoscopic Cholecystectomy

All procedural steps have the same scale:

Did not perform the step	Able to perform a part of the task	Performs the task with much guidance and instructions	Performs the task with minimal guidance and instructions	Can perform the whole task independent, safe and skillful
0	1	2	3	4

i.a. = inapplicable (e.g. because of time shortage)

Step 1. Patient positioning and port insertion

A. Positioning of patient	0	1	2	3	4	i.a.
B. Open introduction	0	1	2	3	4	i.a.
C. Placing of additional trocars	0	1	2	3	4	i.a.

Feedback step 1:

Step 2. Exposure and opening of the peritoneum

A. Placing the patient in reversed Trendelenburg position and tilted to the left	0	1	2	3	4	i.a.
B. Adhesiolysis flush on the gall bladder	0	1	2	3	4	i.a.
C. Exposure of the gall bladder through traction in the right direction with adequate power	0	1	2	3	4	i.a.
D. Opening the peritoneum	0	1	2	3	4	i.a.

Feedback step 2:

Step 3. Dissection of Calot's triangle and achievement of CVS

A. Dissection of fat and fibrous tissue step by step and flush on the gall bladder	0	1	2	3	4	i.a.
B. Exposing the cystic duct and cystic artery at the gall bladder	0	1	2	3	4	i.a.
C. Establishing critical view of safety	0	1	2	3	4	i.a.

Feedback step 3:

Step 4. Clipping and cutting of cystic duct and cystic artery

A. Placing 2 clips central and 1 at the side of the gall bladder	0	1	2	3	4	i.a.
B. Cutting (with cuff > 1 mm)	0	1	2	3	4	i.a.

Feedback step 4:

Step 5. Retrograde/antegrade cholecystectomy

A. Further opening the peritoneum	0	1	2	3	4	i.a.
B. Dissecting the gall bladder from the liver bed	0	1	2	3	4	i.a.
C. Establishing hemostases of the liver bed	0	1	2	3	4	i.a.

Feedback step 5:

Step 6. Ending the operation

A. Using Endobag/Placing pinch over clips	0	1	2	3	4	i.a.
B. Removing the ports under direct vision	0	1	2	3	4	i.a.
C. Closing of fascial defects \geq 5 mm	0	1	2	3	4	i.a.

Feedback step 6:

Figure 1a: Independence-scaled procedural assessment for the laparoscopic cholecystectomy

Feedback:

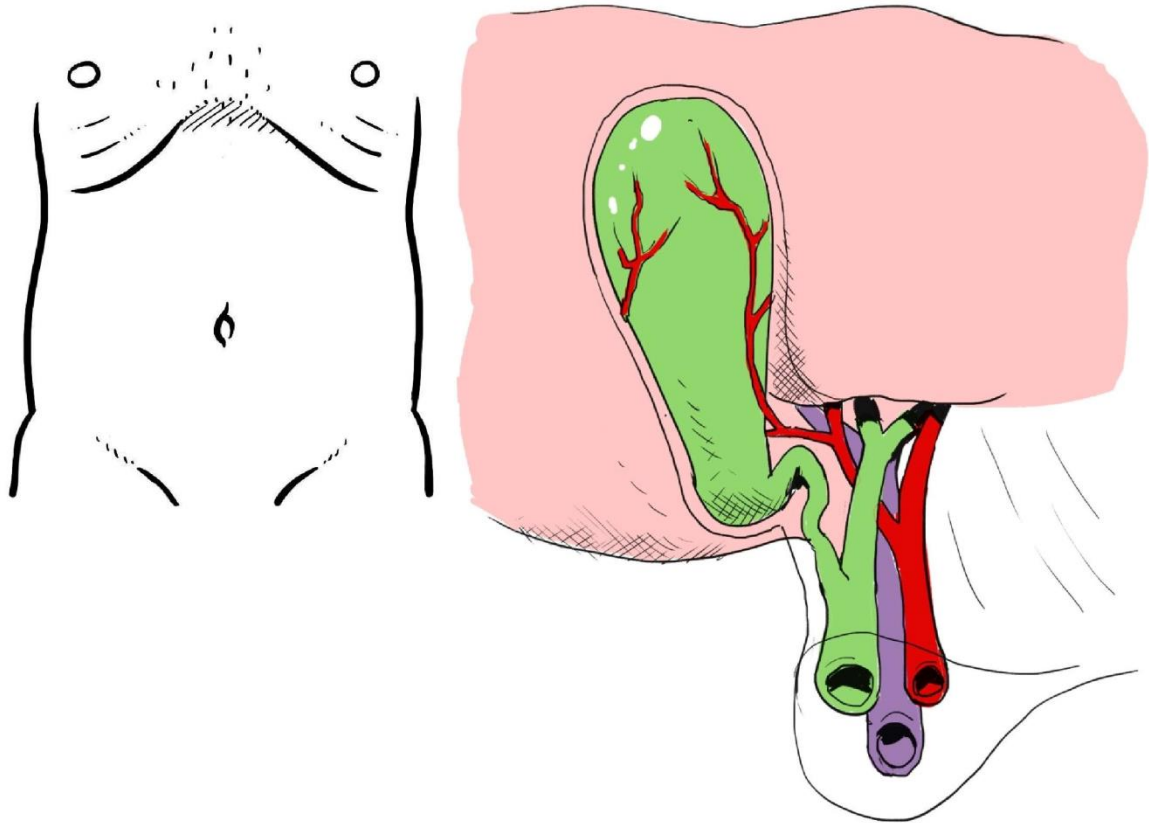


Figure 1b: Independence-scaled procedural assessment for the laparoscopic cholecystectomy

Subjects

To evaluate the validity and reliability of the GRSs and independence-scaled procedural assessment, blinded videos were made and assessed by raters. Videos were made until videos from subjects of three different skill levels were obtained: 1) a novice trainee with prior simulator training, but little experience in the OR (Novice: N = 1-6), 2) an advanced beginner that understands the basic principles, but still has much to learn (Intermediate: N = 7-15) and 3) a trainee that is almost at the point of being qualified to independently perform a procedure, but still operates under direct supervision (Subcompetent: N > 15).

Video recording and blinding

Video and audio recordings were made in the OR with the laparoscope. The communication between the trainee and the supervising surgeon was recorded with 2 tie-pin microphones attached beneath their surgical gown. The recorded audio was used to subtitle the video and to identify the parts in which the supervising surgeon physically assisted or took over a part of the procedure with one or two hands. Verbal communication of the trainee to the supervisor was marked at the beginning of the written sentence with the abbreviation 'Trainee' and of the supervisor to the trainee with the abbreviation 'SV'. Parts performed by the supervisor were made visible in the output video by displaying the abbreviation 'SV-right/left' when the supervisor assisted the procedure with one hand and 'SV' when the supervisor took over with both hands. After subtitling the communication, the videos were muted to prevent voice identification of the trainee and surgeon.

Materials

The communication was recorded with a Shure PG188 PG185 wireless tie-pin microphone (Shure, Culemborg, Gelderland, NL) attached to the trainee and the supervising surgeon beneath their surgical gown. A M-audio M-track USB audio interface (M-audio, Cumberland, Rhode Island, USA) was used in combination with Audacity 2.0.5 software (Free Software Foundation Inc., Boston, USA) to record the transmitted audio on a laptop. Microsoft Windows Moviemaker version 6.0.6000.16386 (Microsoft Corporation, Redmond, USA) was used to synchronize the audio material to the video material, convert the communication to subtitles and mute the video. The final output videos were windows media files of 768x576 pixels, 1000 kb/sec, 4:3 screen ratio and 25 frames/sec. The video material was distributed among raters with USB-sticks in envelopes together with the paper assessment forms randomized in order.

Raters

Ten consultant surgeons and 3 senior surgical trainees (HSTs) from 4 different surgical departments from the North-East Netherlands were invited to participate in the video assessment. In the invitations they were informed that the assessment would take approximately 2.5 hours. The trainees were all in their 4-6th year. In the Netherlands, these are the post graduate training years in which trainees are expected to be able to independently treat uncomplicated gallbladder disease, supervise trainees from the 1-3th year in treating uncomplicated gallbladder disease and perform OSATS assessments of the trainees they have supervised.

Scrub nurses are highly experienced with surgical instruments, but are also familiar with technical requirements of surgeons in the OR. They have seen the total scope of surgical skill levels among trainees and in the majority of cases they possess more OR experience than the operating trainee. Therefore, next to the surgical participants also 6 scrub nurses with working experience in MIS suites were invited to participate in the video assessment.

Assessment instructions, calibration and incentives

In our earlier research with GOALS assessment, we found a relatively low reliability compared to other studies.¹¹ We hypothesized that the lack of exposure and/or training to the assessment method might be one of the contributing factors, as was seen in a series of other studies.^{6,12,13} In this study, the video assessments were therefore preceded by an introduction in order to calibrate the

raters in the following way: 1) The items on the assessment forms were explained, 2) Raters were encouraged to use the full scales as much as possible, 3) Raters were instructed to use their own opinion when rating with the independence-scaled procedural assessment and 4) We attempted to calibrate the raters by giving a clear definition of the low- and high-end of the scale of the GRSs items with a 2 minute operative videos of a novice (N = 1) and of a consultant surgeon (N > 100). We also have hypothesized in the same study that a lack of motivation to complete a comprehensive assessment might lead to unreliable measurements.¹¹ Therefore, those who completed the assessments were rewarded with a box of wine of around 85\$.

Support for implementation

To evaluate the support for implementation of the OSATS, GOALS and independence-scaled procedural assessment among the surgeons and HSTs 6 questions were proposed (Table 1). Five questions could be answered with a score between 1 and 5, with 1 = strongly disagree and 5 = strongly agree. In the 6th question raters were asked whether they rated the assessment tool as a subjective or objective assessment method with 1 = subjective and 5 = objective.

Table 1: Questionnaire about OSATS, GOALS and procedural assessment

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Gives a correct judgment about the competence to perform a specific procedure	1	2	3	4	5
Leads to an unnecessary administrative burden	1	2	3	4	5
Should be used in clinical practice	1	2	3	4	5
Helps in the acquirement of procedural knowledge and skills	1	2	3	4	5
Should also be made for other laparoscopic procedures	1	2	3	4	5
Is objective or subjective	Subjective	Between neutral en subjective	Neutral	Between neutral en objective	Objective

Statistical analysis

Standardized scores

To be able to compare the different assessment methods and to correct for the missing items in GRS ratings and missing and inapplicable items in the independence-scaled procedural assessment score ratings, the ratings were calculated into a standardized percentage score with the formulas:

1. *Procedural assessment score* = $[total\ score / (max.\ score - 4 \times N_{inapplicable} - 4 \times N_{missing})] \times 100$
2. *GRS score* = $[(total\ score - (min.\ score - N_{missing})) / (max.\ score - (min.\ score - N_{missing}) - 5 \times N_{missing})] \times 100$

In the independence-scaled procedural assessment, the items 'positioning of patient', 'open introduction' and 'closing of wounds' were not assessed because they were not captured on the video images of the laparoscopic camera.

Validity

Validity of the assessment tools was estimated by evaluating whether the increase in experience level between trainees in the videos led to a significant increase in performance score with the Friedman's Two-way Analysis of Variance by Ranks. If a significant difference was observed between the video scores, the scores of video 1 and 2 and the scores of video 2 and 3 were compared with the Wilcoxon signed-rank test.

Reliability

The reliability of an assessment tool is dependent on the amount of agreement between ratings of different raters and of crucial importance in high stakes examinations. The reliability was calculated with the ICC (Intra-Class Correlation coefficient). For a detailed discussion of different models to calculate the ICC we refer to the publications of Shrout & Fleiss, McGraw & Wong and Hallgren.¹⁴⁻¹⁶ In this study the absolute agreement 2-way random-effects model for single measures (AA-ICC 2,1) and the consistency agreement 2-way mixed-effects model for single measures (CA-ICC 3,1) of the ICC were chosen. The values that are used to classify the ICC are random in nature and should be adapted to the purpose of the measurement instrument. To evaluate the assessment methods for the purpose of summative assessment a cut-off value of 0.8 was used for the total score of the assessment method.^{4,17} For interpretation of the reliability of the individual items the following cut-off values were used: 'moderate' (0.21-0.40), 'reasonable' (0.41-0.60), 'good' (0.61-0.80) and 'almost perfect' (0.81-1.00).

Feasibility

In the evaluation of feasibility, the assessment methods were compared with the Friedman test. If a statistical significant difference was observed the assessment methods were mutually compared with the Wilcoxon signed-rank test. All statistical analysis were performed with SPSS 20.0.0.1 (SPSS, Chicago, IL, USA). In all analyses a p-value of < 0.05 (2-sided) was considered statistical significant. The Holm-Bonferroni method was applied to correct α for family wise error in the case of multiple testing.

Results

Raters

The surgeons and HSTs (group A) had performed a minimum of 50 LCs and the scrub nurses (group B) had assisted a minimum of 50 LCs. Three surgeons were excluded in group A: 2 surgeons could not participate in the assessment because of time shortage and 1 rater was excluded because 4 of the 9 assessment forms were filled in with identical scores on all items, indicating an incomprehensive assessment. In the residual ratings the maximum number of assessment forms with identical scores on all items was 2.

Videos

Three videos that met the assessment requirements were synchronized, subtitled and blinded. The number of LCs performed, year of training and OSATS score of trainees of the videos are shown in table 2. No significant difference in level of difficulty was observed between the 3 videos ($p = 0.879$, Friedman test).

Validity

Boxplots of the scores of group A and B are shown in Figure 2. In group A, the median OSATS score was 12.5 [0.0-39.3] for video 1, 53.6 [39.3-85.7] for video 2 and 71.4 [50.0-100.0] for video 3 ($p = 0.001$). A significant difference was observed between video 1 and 2 ($p = 0.005$), but not between video 2 and 3 ($p = 0.083$). The median GOALS score was 12.5 [0.0-35.0] for video 1, 53.8 [35.0-90.0] for video 2 and 72.5 [35.0-100.0] for video 3 ($p < 0.001$). A significant difference was observed between video 1 and 2 ($p = 0.005$), but not between video 2 and 3 ($p = 0.096$). The median procedural assessment score was 22.4 [18.3-62.5] for video 1, 65.6 [52.5-91.7] for video 2 and 85.4 [63.5-98.2] for video 3 ($p < 0.001$). In contrast to the GRSs, a significant difference was observed between video 1 and 2 ($p = 0.005$) and between video 2 and 3 ($p = 0.005$).

In group B, the median OSATS score was 9.8 [0.0-28.6] for video 1, 74.1 [50.0-91.1] for video 2 and 83.9 [75.0-98.2] for video 3 ($p = 0.006$). No significant difference was observed between video 1 and 2 ($p = 0.028$) and video 2 and 3 ($p = 0.115$). The median GOALS score was 15.0 [0.0-37.5] for video 1, 66.3 [45.0-90.0] for video 2 and 77.5 [70.0-90.0] for video 3 ($p = 0.009$). No significant difference was observed between video 1 and 2 ($p = 0.027$) and between video 2 and 3 ($p = 0.293$). The median procedural assessment score was 21.7 [11.7-32.1] for video 1, 59.2 [50.0-81.3] for video 2 and 73.8 [59.6-86.5] for video 3 ($p = 0.009$). No significant difference was observed between video 1 and 2 ($p = 0.028$) and between video 2 and 3 ($p = 0.173$).

The median scores of the OSATS, GOALS and independence-scaled procedural assessment items of group A are given in table 3-5. In independence-scaled procedural assessment scores, the scores for video 2 in step 4 'clipping and transection of the cysticus and artery' were excluded, because the cystic duct was too large to be clipped with a clip of normal size. A significant difference between video 1 and 2 and video 2 and 3 was only observed in OSATS item 2 'time and motion'.

Table 3: Standardized score and range of OSATS items for video 1-3 of group A. P-values were calculated with the Friedman test and differences between video 1 and 2 and video 2 and 3 were evaluated with the Wilcoxon test. The Holm-Bonferroni method was applied to correct the significance level. * = Statistical significant.

OSATS						
	Video 1	Video 2	Video 3	p(1-2-3)	p(1-2)	p(2-3)
1. Respect for Tissue	2.0 [1.0-4.0]	3.0 [2.0-5.0]	4.0 [2.0-5.0]	0.002*	0.007*	0.666
2. Time and Motion	1.5 [1.0-3.0]	3.0 [2.0-4.0]	3.5 [3.0-5.0]	<0.001*	0.007*	0.025*
3. Instrument Handling	1.0 [1.0-3.0]	3.0 [3.0-5.0]	4.0 [3.0-5.0]	<0.001*	0.004*	0.305
4. Knowledge of Instruments	2.0 [2.0-3.0]	3.5 [3.0-5.0]	4.5 [3.0-5.0]	0.001*	0.011*	0.084
5. Use of Assistants	1.0 [1.0-4.0]	3.0 [2.0-4.0]	3.5 [2.0-5.0]	<0.001*	0.006*	0.035
6. Flow of Operation	1.0 [1.0-2.0]	3.0 [1.0-4.0]	4.0 [2.0-5.0]	0.001*	0.008*	0.058
7. Knowledge of Procedure	2.0 [1.0-3.5]	4.0 [3.0-5.0]	4.0 [3.0-5.0]	<0.001*	0.005*	0.194

Table 4: Standardized score and range of GOALS items for video 1-3 of group A. P-values were calculated with the Friedman test and differences between video 1 and 2 and video 2 and 3 were evaluated with the Wilcoxon test. The Holm-Bonferroni method was applied to correct the significance level. * = Statistical significant.

GOALS						
	Video 1	Video 2	Video 3	P(1-2-3)	p(1-2)	p(2-3)
1. Depth Perception	1.0 [1.0-4.0]	3.0 [2.0-5.0]	4.0 [2.0-5.0]	0.005*	0.007*	0.589
2. Bimanual Dexterity	2.0 [1.0-2.0]	3.5 [2.0-5.0]	4.0 [3.0-5.0]	<0.001*	0.007*	0.058
3. Efficiency	1.0 [1.0-2.0]	3.0 [3.0-4.0]	4.0 [3.0-5.0]	<0.001*	0.004*	0.096
4. Tissue Handling	2.0 [1.0-3.0]	3.0 [2.0-5.0]	4.0 [2.0-5.0]	0.005*	0.017*	0.341
5. Autonomy	1.0 [1.0-2.0]	2.5 [1.0-4.0]	4.0 [1.0-5.0]	0.001*	0.007*	0.047

Table 5: Standardized score and range of procedural assessment items for video 1-3 of group A. P-values were calculated with the Friedman test and differences between video 1 and 2 and video 2 and 3 were evaluated with the Wilcoxon test. The Holm-Bonferroni method was applied to correct the significance level. * = Statistical significant.

Independence-scaled procedural assessment						
	Video 1	Video 2	Video 3	P(1-2-3)	p(1-2)	p(2-3)
1 Positioning and introduction of the trocars	25.0 [0.0-75.0]	75.0 [50.0-100.0]	87.5 [75.0-100.0]	<0.001*	0.007*	0.096
2 Exposition gallbladder and opening of peritoneum	33.3 [18.8-43.8]	75.0 [41.7-100.0]	91.7 [66.7-100.0]	<0.001*	0.005*	0.042
3 Dissection of Calot's triangle	12.5 [0.0-66.7]	43.8 [25.0-75.0]	66.7 [25.0-91.7]	<0.001*	0.005*	0.192
4 Clipping and transection of the cysticus and artery	12.5 [12.5-75.0]		100.0 [75.0-100.0]	-	0.004*	
5 Retrograde/antegrade cholecystectomy	29.2 [16.7-75.0]	75.0 [33.3-100.0]	100.0 [75.0-100.0]	<0.001*	0.011*	0.026
6 Extraction of gallbladder and closing of wounds	25.0 [0.0-50.0]	75.0 [75.0-100.0]	93.8 [75.0-100.0]	<0.001*	0.005*	0.482

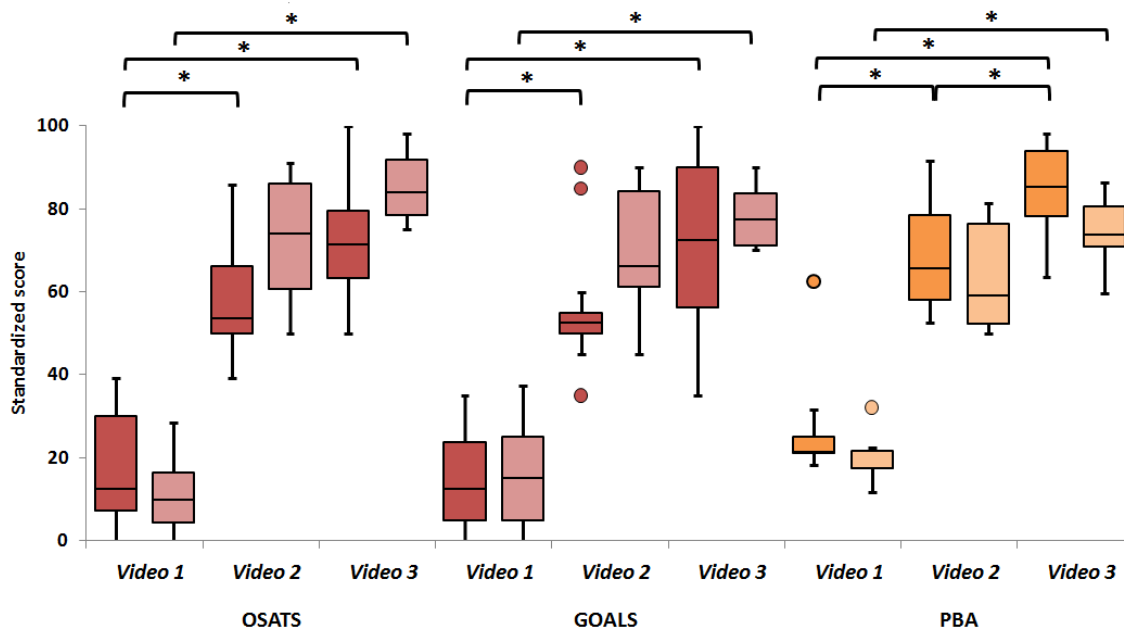


Figure 2: Standardized scores for video 1, 2 and 3 (novice, intermediate and subcompetent) for independence-scaled procedural assessment and global rating scales (OSATS and GOALS). Procedural assessment and global rating scales scores (GRS) improved significantly with surgical experience (OSATS $p < 0.001$, GOALS $p < 0.001$, PBA $p < 0.001$). However, the PBA was the only one of the three assessment methods that could differentiate between the video of the intermediate and sub competent trainee among the surgical raters ($p = 0.005$).

Reliability

The AA-ICC and CA-ICC of the OSATS, GOALS and independence-scaled procedural assessment scores and their individual items are shown in table 6-8. The AA-ICC of the total OSATS score was 0.78 in group A and 0.91 in group B. Most OSATS items had a good or almost perfect reliability in both groups, except for the items respect for tissue and use of assistance. Interestingly, the two items 'use of assistance' and 'instrument handling' attained an AA-ICC and CA-ICC of ≥ 0.90 in group B.

The AA-ICC of the total GOALS score was 0.74 in group A and 0.85 in group B. The AA-ICC and CA-ICC of the items 'depth perception' and 'tissue handling' were reasonable in group A.

The AA-ICC of the total independence-scaled procedural assessment score was 0.84 in group A and 0.87 in group B. The procedural step dissection of Calot's triangle had a reasonable ICC, only the CA-ICC in group A was good.

When group B was added to group A, the ICCs of the total scores and items were higher than that of group A in all 3 assessment methods, except for dissection of Calot's triangle.

Table 6: AA-ICC 2,1 and CA-ICC 3,1 of standardized total OSATS score and the standardized score of the items of the OSATS. All ICCs were statistical significant ($p < 0.05$).

Item	Group A + B		Group A		Group B	
	AA-ICC	CA-ICC	AA-ICC	CA-ICC	AA-ICC	CA-ICC
1. Respect for Tissue	0.50	0.51	0.47	0.49	0.49	0.46
2. Time and Motion	0.71	0.76	0.71	0.77	0.74	0.75
3. Instrument Handling	0.78	0.80	0.71	0.70	0.90	0.94
4. Knowledge of Instruments	0.76	0.79	0.72	0.72	0.82	0.90
5. Use of Assistants	0.70	0.80	0.58	0.74	0.90	0.92
6. Flow of Operation	0.74	0.77	0.64	0.68	0.88	0.89
7. Knowledge of Procedure	0.76	0.73	0.66	0.65	0.86	0.83
Total	0.83	0.84	0.78	0.79	0.91	0.92

Table 7: AA-ICC 2,1 and CA-ICC 3,1 of standardized total GOALS score and the standardized score of the items of the GOALS. All ICCs were statistical significant ($p < 0.05$).

Item	Group A + B		Group A		Group B	
	AA-ICC	CA-ICC	AA-ICC	CA-ICC	AA-ICC	CA-ICC
1. Depth Perception	0.64	0.71	0.49	0.53	0.84	0.95
2. Bimanual Dexterity	0.78	0.83	0.76	0.78	0.83	0.90
3. Efficiency	0.86	0.89	0.87	0.87	0.84	0.91
4. Tissue Handling	0.56	0.56	0.49	0.46	0.59	0.64
5. Autonomy	0.66	0.69	0.60	0.67	0.78	0.72
6. Level of Difficulty	NS	NS	NS	NS	NS	NS
Total	0.79	0.81	0.74	0.75	0.85	0.89

Table 8: AA-ICC 2,1 and CA-ICC 3,1 of standardized total procedural assessment score and the standardized score of the items of the procedural assessment. In step 1, 'positioning' (= pre-operative positioning) was not assessed and in step 6 'closing of wounds' was not assessed. All ICCs were statistical significant ($p < 0.05$).

Procedural step	Group A + B		Group A		Group B	
	AA-ICC	CA-ICC	AA-ICC	CA-ICC	AA-ICC	CA-ICC
1. Positioning and introduction of the trocars	0.82	0.80	0.79	0.77	0.89	0.86
2. Exposition gallbladder and opening of peritoneum	0.79	0.76	0.82	0.80	0.71	0.66
3. Dissection of Calot's triangle	0.45	0.59	0.50	0.63	0.52	0.52
4. Clipping and transection of the cysticus and artery	0.92	0.94	0.90	0.92	0.97	0.97
5. Retrograde/antegrade cholecystectomy	0.74	0.75	0.75	0.74	0.67	0.71
6. Extraction of gallbladder and closing of wounds	0.89	0.88	0.86	0.84	0.92	0.92
Total	0.85	0.87	0.84	0.86	0.87	0.86

Support for implementation

Seven surgeons and three surgical trainees completed the questionnaire (Figure 3). All shared the opinion that the independence-scaled procedural assessment score gives a correct judgment of competency in a specific procedure, compared to 6 for the OSATS and 4 for the GOALS ($p = 0.001$). A significant difference was observed between the independence-scaled procedural assessment and the GRSs ($p = 0.011$ for OSATS, $p = 0.005$ for GOALS). Four raters found the independence-scaled procedural assessment an unnecessary administrative burden, compared to 4 for the OSATS and 2 for the GOALS ($p = 0.692$). They all thought that the independence-scaled procedural assessment should be used in clinical practice, compared to 2 for the OSATS and 3 for the GOALS ($p = 0.005$). A significant difference was observed between the independence-scaled procedural assessment and the GRSs ($p = 0.018$ for OSATS, $p = 0.010$ for GOALS). Six raters agreed on the statement that the independence-scaled procedural assessment could help in the acquirement of procedural knowledge and skills compared to 2 for the OSATS and 2 (2 out of 9 because of missing data from 1 rater) for the GOALS ($p = 0.025$). A significant difference was only observed between the independence-scaled procedural assessment and the OSATS in this question ($p = 0.009$). Eight observers considered the independence-scaled procedural assessment to be objective compared to 3 for the OSATS and 3 for the GOALS ($p = 0.007$). A significant difference was observed between the independence-scaled procedural assessment and the GRSs ($p = 0.015$ for OSATS, $p = 0.023$ for GOALS). All participants encouraged a reproduction of the independence-scaled procedural assessment for other laparoscopic procedures.

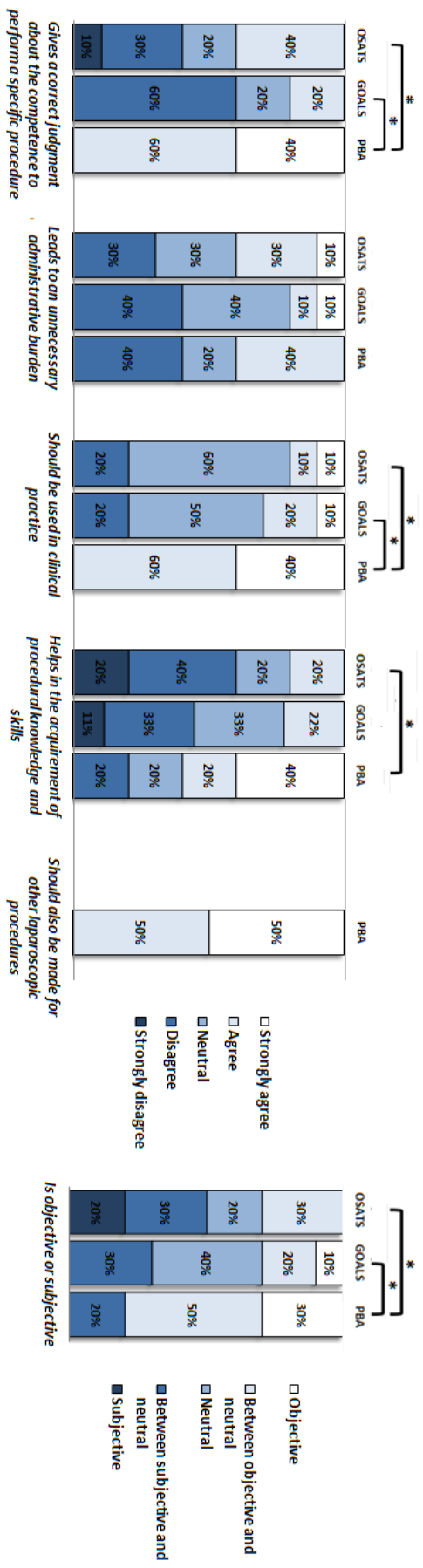


Figure 3: Support for implementation of competence-based assessment.

Discussion

Although GRSs have proven its value in formative feedback in training, controversy exists about their usefulness in procedure specific assessment and certification for independent surgical treatment of uncomplicated disease. A multicenter blinded study was conducted to estimate the validity, reliability and feasibility of the procedural assessment and two GRS of which one, the OSATS, is an integral part of surgical training in the Netherlands. A procedural assessment for the LC was created by linking the previously published operative key steps to an independence scale to create a procedural assessment.⁹ Three blinded and subtitled videos of trainees of different skill levels were assessed with the independence-scaled procedural assessment, OSATS and GOALS by surgeons, senior surgical trainees and scrub nurses. In addition, a questionnaire was distributed that aimed to measure the support for implementation of the independence-scaled procedural assessment, OSATS and GOALS in practice.

Validity

The independence-scaled procedural assessment, OSATS and GOALS all showed a significant improvement in assessment scores with increasing experience levels. This supports the results of previous studies that have evaluated the validity of GRSs and independence-based procedural assessment.^{3,6,10,11} However, in this study the independence-scaled procedural assessment was the only one of the three assessment methods that could differentiate between the video of the intermediate and subcompetent trainee among the surgical raters. This indicates that the independence-based procedural assessment is the most sensitive assessment method to measure skill level in the performance of a procedure and is in line with recent studies that studied independence scales. For instance, Glarner et al. used an independence scale as an indirect measure of the skill level of the surgeon for assessment of a hemicolectomy.¹⁰ Their independence-scaled procedural assessment showed an increase in performance level in residents during a colorectal rotation, while the GRSs showed little to no increase during the rotation. Cornelis et al. have shown that the so-called 'Alphabetic Summary Scale', an independence-based rating scale, had a higher discriminating power than a modified form of the OSATS and an overall performance scale for assessment of osteosynthesis of proximal femoral fractures.³

Next to the higher sensitivity, the independence-scaled procedural assessment also has the advantage of providing educators and trainees with the opportunity to preoperatively discuss which procedural steps will be performed by the trainee and assessed by the supervisor. This enables a stepwise expansion of the amount of steps performed by a trainee. GRSs lack the benefits of enabling stepwise teaching and the use of solely a GRS to assess operative competence and therefore probably do not optimally facilitates the teaching of procedural skills. The GRSs also lack an option for narrative (descriptive) feedback. We decided to include multiple options for giving narrative feedback in the independence-scaled procedural assessment, which makes it more suitable for giving feedback that is task-specific and focused on the learning goals of a trainee.¹⁷

Reliability

This is the first blinded multicenter study that simultaneously investigates the reliability of GRSs and independence-based procedural assessment for a standard laparoscopic procedure. The patterns observed in the reliability analysis give valuable insights in the factors that influence reliability in the assessment of surgical competence.

Among the raters with surgical training, the reliability of the GRSs did not reach the threshold of 0.8. This finding is in line with the majority of studies that addressed the reliability of GRSs.⁴ There are a series of factors that could have led to an inter-rater reliability below the threshold value. In the past, authors have argued that training might be of key importance in attaining reliable scores with GRSs.^{6,11,12} Because the OSATS is an integral part of surgical training in the Netherlands, all surgical raters were familiar with this assessment method. However, some of the raters had never used the

other 2 assessment methods to assess operative competence. We attempted to introduce raters to the key elements of the assessment methods and to calibrate them with short introductory videos prior to assessment. In both GRSs, the introduction and calibration did not lead to an acceptable reliability for summative assessment.

Assuming the introduction to assessment was done appropriately, the most likely remaining cause of not attaining the threshold are characteristics of the GRSs itself. The format of the GRSs, in particular the Likert scale, has been subject of discussion. Some authors even state that attaining a reliability of 0.80 is almost impossible when using a Likert scale.¹⁸ The descriptions of the anchors show a possible weakness of the GRSs. The anchors contain words such as 'frequently', 'unnecessary' and 'inappropriate', that are strongly susceptible to differences in interpretation and the absence of descriptions on anchors with score 2 and 4 might increase subjectivity even more. The terminology and characteristics of the scale probably contribute to a barrier for attaining a high inter-rater reliability with GRSs.

In contrast to the GRSs, the independence-scaled procedural assessment showed an inter-rater reliability higher than 0.8 among surgeons, indicating that an independence-based procedural assessment tool is a suitable candidate for certification and authorization in the treatment of uncomplicated disease. This is in line with the observation of an ICC higher than 0.8 by Miskovic et al., who evaluated independence-scaled procedural assessment in colorectal surgery and determined inter-rater reliability by correlating peer- with self-assessments.¹⁹ It seems that assessment of a series of procedural key steps, on which consensus has been achieved, compels raters to look at specific elements of operative competence and thereby gives less room for subjectivity. The high inter-rater reliability could theoretically also have been caused by a higher between-subjects variance in the independence-scaled procedural assessment: if the performance level of trainees with different experience levels measured with a procedural assessment shows more variance than when assessed with a global assessment method, the reliability of the former would automatically tend to increase based on the calculation model of the ICC.²⁰ However, comparison of the between-subjects mean square of the independence-scaled procedural assessment and GRSs did not indicate that this was the case.

Although the total independence-scaled procedural assessment scores showed a high reliability, subjectivity was not totally expelled. This was especially evident in the inter-rater reliability of the dissection of Calot's triangle. Interestingly, among surgeons the CA-ICC was good, indicating that part of the error variance is caused by some clinical supervisors being more stringent than others in the assessment of this step. To increase the inter-rater reliability in this procedural step, a more detailed procedure characterisation with the inclusion of procedure errors could have been included as has been done by others.^{21,22} However, several researchers in the domain of performance appraisal have proposed an alternative view on inter-rater reliability that might be relevant in the assessment of the dissection of Calot's triangle. This view has been described by Govaerts et al. as the 'constructivist social-psychological approach'.²³ One of the central themes of this perspective is that "raters from different perspectives may rate differently because they observe different aspects of performance, and differences in ratings may very well reflect true differences in performance." The dissection of Calot's triangle is the most complex and therefore the most technically demanding step. Because the high difficulty requires a mixture of technical behaviors in the trainee, the rater has to make a decision on which aspect of technical behavior of the trainee to rate during the observation of the behavior during this step and also has to decide on which way it will be assessed. These decision processes are influenced by knowledge, operative experiences and the content and characteristics of the interactions with supervisors who supervised the rater (socialization). Thus, although the ratings do not agree in the assessment of the dissection of Calot's triangle, they might all be equally valid, because they are founded on the individual professional experience and understanding of the raters. If so, this could have the implication that a summative assessment of a trainee would not be based on the assessment of one rater, but on multiple raters, not to achieve a more reliable numerical score, but to achieve a more complete picture of the level of surgical skills.²³ For instance, a trainee would only be considered eligible for certification in the

independent treatment of uncomplicated gallbladder disease if a specific cut-off score is achieved on two laparoscopic cholecystectomies, each supervised by a different consultant surgeon.

At last, when the ratings of the scrub nurses were combined with those of the surgically trained raters, almost all the reliability coefficients of the total scores and item scores increased slightly, indicating that, in line with the study of Beard et al, there is agreement between the assessment of scrub nurses and surgeons.²⁴ Although the authorization of surgical trainees in the independent treatment of patients with uncomplicated disease should be reserved for clinical supervisors, these findings indicate that scrub nurses can be of contributive value in the assessment of operative competence of trainees.

Support for implementation

In the questionnaire, there was strong support for implementation of the independence-scaled procedural assessment into practice. Although we did not give an extensive description on what can go good and what can go wrong, it was considered to give a more correct judgment of procedural skills than the GRSs. Participants were also asked to rate the assessment methods on objectivity. The median score of objectivity for the OSATS and for the GOALS in this study was 2.5 and 3.0 resp., which is similar to the median score of 3.0 observed by Hiemstra et al. on the same question for the OSATS among gynecologists and gynecological residents.⁵ However, eight out of ten considered the independence-scaled procedural assessment to be objective (median score = 4.0). Furthermore, all participants encouraged reproduction of the independence-scaled procedural assessment for other laparoscopic procedures. These findings are in line with the findings of Beard et al. who have shown a higher acceptability and satisfaction of their procedure based assessment than for the OSATS among trainees and clinical supervisors.⁸

Development of procedural based assessments

On the basis of the results we recommend using a two step system for the development for procedural assessments (Figure 4). The first step consists of using a regional expert panel to reach consensus on the key steps of a procedure. The procedural steps that are considered of key importance in a procedure can vary regionally and internationally. By using the opinion of experienced surgeons involved in surgical training programs within the region, the procedural steps will be relevant and important to those using it (content validity). In the second step an independence scale is attached to the key steps to assess operative competence.

An alternative to the second step would be to give elaborate descriptive terms of how the key steps of a procedure should be performed or to insert some form of error analysis in the assessment as has been done by others.^{21-22, 25-30} However, error-based assessment might be limited in assessment above the performance level of what Wentink et al. call skill and rule-based behaviour.³¹ The higher levels of cognition, by Wentink et al. described as 'knowledge based behaviour', are used for the development and execution of strategies to deal with unfamiliar situations during surgery.³¹ This level of behavior moves more to the foreground in the last part of the learning curve, the phase in which skill and rule-based behavior have been largely acquired, but reasoning might need some important adjustments at times. The independence-scaled assessment method gives supervisors the freedom of assessing the level of knowledge-based behavior on the basis of their professional judgment of unfamiliar situations and the adequacy of the trainee's response on these situations. This aspect of assessment is essential in identifying trainees who are ready for independent surgical treatment of patients. Future studies that compare independence-based procedural assessment, error-based procedural assessment and checklist-based procedural assessment in terms of validity, reliability and feasibility could provide more insight on the strengths and weaknesses of each of these assessment methodologies.

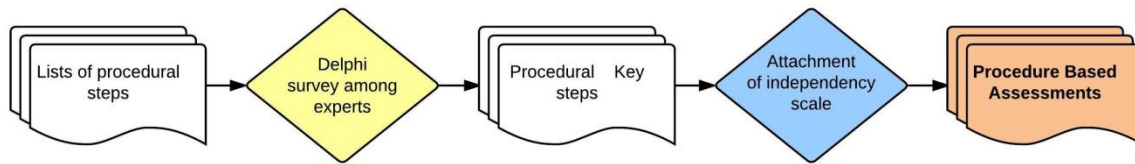


Figure 4: Proposed two-step method for the development of procedure based assessment forms.

Limitations

There are some limitations to our study that have to be addressed. First, the videos were blinded but not randomized. Not using a random sequence could have introduced bias in the assessment. However, as some raters rated video 3 lower than video 2, we do not think that not randomizing the videos affected the raters significantly.

Second, the error variance could have been lower in the independence-scaled procedural assessment because the raters simply did not use their own opinion but adopted that of the supervising surgeon of the video, resulting in a higher reliability than the GRSs. The scrub nurses might be particularly susceptible to this, but the reliability of the independence-scaled procedural assessment of the scrub nurses was similar to that of the GRSs. Therefore, there is no indication that this phenomenon might have artificially increased the reliability of the independence-scaled procedural assessment.

Third, although the literature agrees about using 0.80 as a threshold when assessing reliability for high-stakes examinations, the use of a somewhat arbitrary number as a threshold is arguable. A threshold of 0.80 only means that 80% of the difference between ratings is attributable to true variance and the remaining is caused by random error, rater error and/or other sources of error. Despite this weakness, the threshold is one of the few tools available to identify assessment methods with an inter-rater reliability satisfactory for summative assessment and is strongly adhered to in the surgical literature.⁴

Fourth, no attempts were made to define cut-off values for the independent surgical treatment of uncomplicated gallbladder disease. Research is currently being conducted in our center to collect the required data to establish cut-off values for the identification of competent trainees.

Fifth, after the achievement of a certain skill level, a decay effect has been observed of the acquired skills.³²⁻³⁴ The amount of decay that arises is dependent on 2 variables: 1) How familiar the trainee is with the skills and 2) The amount of time that has passed since the last performance. Although we expect that the independence-scaled procedural assessment is able to identify the level of procedural skills required for the LC, no statements can be made about the number of procedures that have to be performed in order to minimize the decay effect or the length of time the acquired level of procedural skills will be retained. It could furthermore be that the rather verbal passive form of training necessary for adequate independence-scaled procedural formative assessment, increases the retention of skills as described by the guidance hypothesis.^{35,36}

Finally, assessment of non-technical skills such as medical knowledge, communication skills and clinical judgment were not included in this study. Non-technical skills are critical components of operative care and should complement assessment of technical skills when surgical competence is addressed.

Conclusion

In conclusion, a valid and reliable procedural assessment method can be developed by linking the key steps of a procedure, composed with the Delphi methodology, to an independence-based scale. The validity and reliability of the independence-scaled procedural assessment exceeded that of the global rating scales in the blinded assessment of a laparoscopic cholecystectomy. Among the group of raters with surgical training an inter-rater reliability above the threshold value of 0.80 was only observed in the procedural assessment. Moreover, the participants expressed strong support for the use of the independence-scaled procedural assessment in clinical practice and encouraged its reproduction for other procedures. This study demonstrates that independence-scaled procedural assessment is a valuable assessment tool and appears to comply with the requirements of use for procedural certification.

Acknowledgements

We would like to thank Christiaan Hoff, Erik Totte, Jan Apers, Koen Talsma, Renske Schasfoort, Patrick Hemmer, Johan Lange, Carlijn Buis, Frederieke Dijkstra, Mirjam Kaijser, Robbert Bosker, Ilona Pereboom, Kevin Wevers, Ingeborg Riedstra, Wiep Rienks, Linda van de Meulen, Jeroen Kindt, Hindrik Boonstra, Fronnie Kramer, Gerda Kootstra, Lotte van der Werff and the surgical trainees at the Medical Center Leeuwarden for their dedicated participation in our study.

References

- 1 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;84:273–278.
- 2 Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, *et al.* Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today.* 2013;43:271–275.
- 3 Hopmans CJ, den Hoed PT, van der Laan L, van der Harst E, van der Elst M, Mannaerts GHH, *et al.* Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study. *Surgery.* 2014;156:1078–1088.
- 4 Van Hove PD, Tuijthof GJ, Verdaasdonk EGG, Stassen LP, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg.* 2010;97:972–987.
- 5 Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. *Can J Surg.* 2011;54:116–122.
- 6 Vassiliou MC, Feldman LS, Fraser SA, Charlebois P, Chaudhury P, Stanbridge DD, *et al.* Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov.* 2007;14:211–216.
- 7 Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, *et al.* A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107–113.
- 8 Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Health Technol Assess.* 2011;15:161–162.
- 9 Bethlehem MS, Kramp KH, van Det MJ, Ten Cate Hoedemaker HO, Veeger NJ, Pierie JP. Development of a standardized training course for laparoscopic procedures using delphi methodology. *J Surg Educ.* 2014;71:810–816.
- 10 Glarner CE, McDonald RJ, Smith AB, Levenson GE, Peyre S, Pugh CM, *et al.* Utilizing a novel tool for the comprehensive assessment of resident operative performance. *J Surg Educ.* 2013;70:813–820.
- 11 Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJ, Pierie J-PENP. Validity and Reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in Novice Trainees Performing a Laparoscopic Cholecystectomy. *J Surg Educ.* 2015;72:351–358.
- 12 Matsuda T, Kanayama H, Ono Y, Kawauchi A, Mizoguchi H, Nakagawa K, *et al.* Reliability of laparoscopic skills assessment on video: 8-year results of the endoscopic surgical skill qualification system in Japan. *J Endourol.* 2014;28:1374–1378.
- 13 Schijven MP, Reznick RK, ten Cate OT, Grantcharov TP, Regehr G, Satterthwaite L, *et al.* Transatlantic comparison of the competence of surgeons at the start of their professional career. *Br J Surg.* 2010;97:443–449.
- 14 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420–428.
- 15 Hallgren KA. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol.* 2012;8:23–34.
- 16 McGraw K, Wong S. Forming inferences about some intraclass correlation coefficients. *Psychological Methods.* 1996;1:30–46.
- 17 Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. The use of observational diaries in in-training evaluation: student perceptions. *Adv Health Sci Educ Theory Pract.* 2005;10:171–188.

- 18 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;17:1525–1529.
- 19 Miskovic D, Wyles SM, Carter F, Coleman MG, Hanna GB. Development, validation and implementation of a monitoring tool for training in laparoscopic colorectal surgery in the English National Training Program. *Surg Endosc.* 2011;25:1136-1142.
- 20 Foundations of clinical research – Application to practice, 3th edition. Portney LG, Watkins MP. Pearson Education, Inc., New Jersey 07458.
- 21 Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *The American Journal of Surgery*, 2007;193:797-804.
- 22 Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of Surgery*, 2002;236:458-463.
- 23 Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract.* 2007;12:239-260.
- 24 Beard JD. Setting standards for the assessment of operative competence. *Eur J Vasc Endovasc Surg.* 2005;30:215–218.
- 25 Palter VN, MacRae HM, Grantcharov TP. Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. *Am J Surg.* 2011;201:251–259.
- 26 Sarker SK, Chang A, Vincent C, Darzi SA. Development of assessing generic and specific technical skills in laparoscopic surgery. *Am J Surg.* 2006;191:238–244.
- 27 Eubanks TR, Clements RH, Pohl D, Williams N, Schaad DC, Horgan S, et al. An objective scoring system for laparoscopic cholecystectomy. *J Am Coll Surg.* 1999;189:566–574.
- 28 Kurashima Y, Feldman LS, Al-Sabah S, Kaneva PA, Fried GM, Vassiliou MC. A tool for training and evaluation of laparoscopic inguinal hernia repair: the Global Operative Assessment Of Laparoscopic Skills-Groin Hernia (GOALS-GH). *Am J Surg.* 2011;201:54–61.
- 29 Hwang H, Lim J, Kinnaird C, Nagy AG, Panton ON, Hodgson AJ, et al. Correlating motor performance with surgical error in laparoscopic cholecystectomy. *Surg Endosc.* 2006;20:651–655.
- 30 Zevin B, Bonrath EM, Aggarwal R, Dedy NJ, Ahmed N, Grantcharov TP. Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. *J Am Coll Surg.* 2013;216:955–965.
- 31 Wentink M, Stassen LP, Alwayn I, Hosman RJ, Stassen HG. Rasmussen's model of human behavior in laparoscopy training. *SurgEndosc.* 2003;17:1241-1246.
- 32 Maagaard M, Sorensen JL, Oestergaard J, Dalsgaard T, Grantcharov TP, Ottesen BS, et al. Retention of laparoscopic procedural skills acquired on a virtual-reality surgical trainer. *Surg Endosc.* 2011;25:722–727.
- 33 Edelman DA, Mattos MA, Bouwman DL. FLS skill retention (learning) in first year surgery residents. *J Surg Res.* 2010;163:24–28.
- 34 Sinha P, Hogle NJ, Fowler DL. Do the laparoscopic skills of trainees deteriorate over time? *Surg Endosc.* 2008;22:2018–2025.
- 35 Schmidt RA, Bjork RA. New conceptualizations in practice: common principles in three paradigms suggest new concepts for training, *Psychol Sci* 1992;3:207.
- 36 Winstein CJ, Pohl PS, Lewthwaite R. Effects of physical guidance and knowledge of results on motor learning: support for the guidance hypothesis. *Res Q Exerc Sport.* 1994;65:316-323.

Chapter 9
General discussion and future perspectives

General discussion and future perspectives

After the achievement of proficiency criteria in laparoscopic skills training outside the OR, trainees commence their learning experience in the OR supervised by a consultant surgeon who guides them through the procedure. This thesis is focused on the enhancement of these educational efforts. The goal of this thesis is to improve: 1) assessment of candidates for medical specialties that require laparoscopic surgery, 2) procedural laparoscopy training on the OR and 3) post-operative procedure specific performance feedback and assessment.

Part I: Aptitude

The question of whether aptitude tests that evaluate visual-spatial ability, perceptual ability and psychomotor ability can be used in the assessment of candidates for medical specialties that require laparoscopic skills is currently a topic of debate. We conducted a meta-analysis to: 1) evaluate whether aptitude assessments can be used to predict the ability to acquire and perform laparoscopic skills, 2) to quantify how much of the variability in skills can be predicted by aptitude assessment and 3) obtain insight in the factors that influence the strength of this relationship.

Although no statements can be made about surgery in general, we can state that the acquisition and performance of laparoscopic skills can partly be predicted with aptitude measurements. Assessment of aptitude in the form of visual-spatial ability tests, perceptual ability tests, psychomotor ability tests and simulator-based assessment all showed a significant correlation with surgical training. A significant correlation was also found when only studies that used aptitude tests to predict performances during an OR training session were singled out.

The ergonomic challenges of the OR environment encountered during surgical procedures on human beings provide a theoretical support for the association of laparoscopic surgery with the content of these aptitude tests. As the results in the meta-analysis support this construct, program directors can feel legitimized to use a laparoscopy aptitude test (LAT) in the assessment of candidates that require laparoscopic skills, even without the extensive validation of these tests on the basis of job performances as a fully certified laparoscopic surgeon, which is the ultimate measure of predictive validity. The latter is a difficult task considering the fact that during in vivo laparoscopy there can always be unexpected visual-spatial, perceptual or psychomotor challenges that place high demands on the cognitive abilities of laparoscopic surgeons. It might be hard to estimate individual performance level on these instances, as operative demands that exceed the capacity of the surgeon often involve an inversion of the indirect and direct control dynamics described in chapters 3 and 8, manifesting as the cognitive or physical support by another, often more experienced, surgeon in the department. At other times, the increased demands might be only expressed in a longer operation time, a variable which is also determined by a multitude of other variables.

It is important to keep in mind that a LAT can be used to optimize the selection process of candidates, but can be just as beneficial for career coaching of medical students. A LAT can help make the right career decision and/or support surgical educators in the recommendation to opt for a specific area of medicine. Students tested with a high aptitude interested in a non-surgical career can obtain a stimulus to consider pursuing a surgical career and those with a low aptitude interested in a career involving laparoscopy have the opportunity to invest their valuable time and energy in a specialty or differentiation program that better matches their talent or to keep persuing their dream knowing that they might have to work harder to attain the same level of competency as their peers.

If the evidence supports the theory that visual-spatial, perceptual and psychomotor ability are predictors of laparoscopic skills, which aptitude measurement should be used in a LAT? As simulators seem to measure all 3 forms of aptitude at once and are available in most academic hospitals with surgical training programs, the most straightforward option would be to use these devices for aptitude assessment with a norm-referenced scoring system. On the other hand, the increasing

availability of simulators in the form of serious games¹ introduces the danger of measuring the amount of adaptation to a human-machine interface instead of aptitude. Other options which include the use of a test battery of visual-spatial ability, perceptual ability and psychomotor ability, with or without simulator based assessment, are organizational demanding and therefore financially more burdening. Thus, the question is whether we can find a way to prevent adaptability to the human-machine interface to become a problem in aptitude assessment. Perhaps, the most practical solution would be to introduce an instructional course and allow free practice on freely accessible simulators for medical students and subsequently use a set of inaccessible difficult simulator tasks, tasks that have a high level of unpredictability and have the ability to distinguish in aptitude even after motivated in vitro training, to assess these students during a LAT (Figure 1). This idea is supported by a recent publication that showed that the relationship of aptitude with laparoscopic skills remains present also after an instructional course in combination with voluntary practice.²

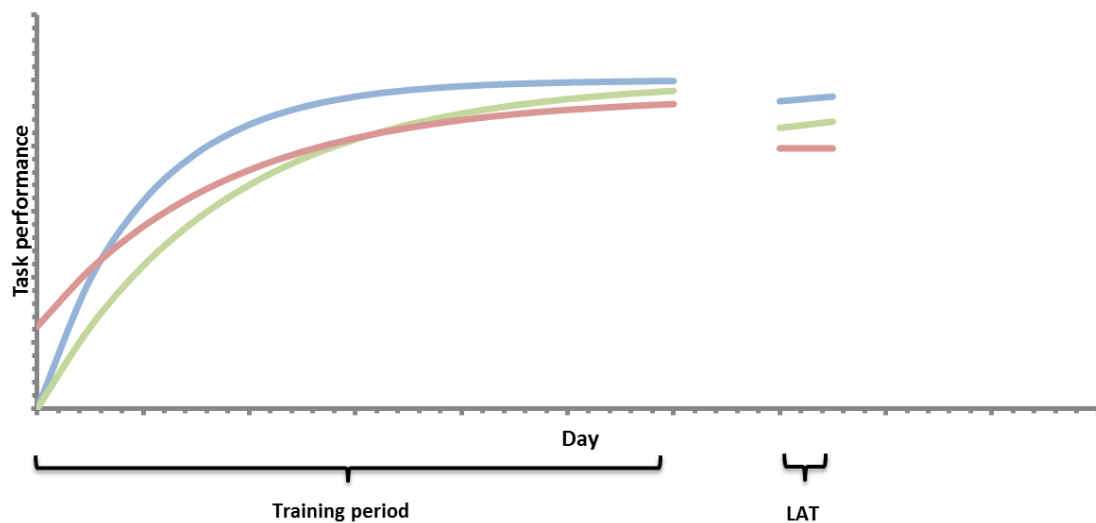


Figure 1: Laparoscopy Aptitude Test (LAT) is preceded by a time period of supervised simulator training and free practice on simulators for applicants. The LAT itself contains unpredictable simulator tasks that are inaccessible to applicants and has the ability to distinguish different levels of cognitive ability in motivated simulator trained subjects. Red: applicant with previous experience on serious video game but low aptitude score on LAT, blue: applicant with no previous experience but high aptitude score on LAT, green: applicant with no previous experience and intermediate aptitude score on LAT.

Part II: Training

In 1887, Vilfredo Pareto published his observation of an exponential law between the amount of wealth an inhabitant owned and the rank-order of the inhabitant.³ He concluded that 80% of property is owned by merely 20% of the inhabitants. This concept, that a relative few account for a large proportion of a common effect, also has become known as the '80-20 rule'. We used the Pareto-analysis to state content criteria for surgical training. On the basis of the results, it seems that the Pareto-analysis is a tool with high potential for identifying on-the-job challenges as 35 of 253 (13.8%) of the different verbal corrections were responsible for 80% of the total number of verbal corrections counted. We have suggested training methods with a high content validity on the basis of the results. These methods vary from simple instructional courses to technically complex gadgets to enhance training efficiency. Further research is necessary to evaluate the effectiveness of surgical education methods developed on the basis of a Pareto-analysis.

To investigate the different positions that can be used to perform a laparoscopic cholecystectomy on patients, we conducted a cross-over study of the French versus the American position. No statistically significant difference was found between the French and American position in the

posture of the vertebral column among surgeons. On the basis of the place of trocar insertion however, it can be hypothesized that the left arm is at risk for being overburdened in the American position, the position most often used in the Netherlands. Research focused on the shoulders, arms and hand movements instead of the vertebral column might be able to provide more information about the degree of strain of the upper extremities in the American and the French position, hence, making the preference for one of the two operation setups in surgical education more justifiable.

Besides the attempt to reach consensus on a preferable operation position for the laparoscopic cholecystectomy, a Delphi survey was performed to reach consensus on the key steps of 2 basic laparoscopic procedures, the laparoscopic cholecystectomy and the laparoscopic appendectomy. The Delphi method has recently been used by our research group to establish the key steps of more sophisticated laparoscopic procedures: the laparoscopic right hemicolectomy and the laparoscopic sigmoid colectomy.⁴ We consider reaching consensus on key steps of laparoscopic procedures as the first steps towards a standardized curriculum in laparoscopic surgery training. Delphi survey based key steps can facilitate deliberate practice on the following ways:

1. Pre-operative preparation for OR training.
2. A roadmap for the stepwise teaching in surgical training for laparoscopic procedures.
3. Post-operative assessment of procedural skills.
4. Certification for the surgical treatment of uncomplicated disease.

In part III we evaluate whether the key steps of the laparoscopic cholecystectomy can be used for two of these goals, the post-operative assessment of procedural skills and the examination for independent surgical treatment.

Part III: Assessment

Assessment of procedural learning is an important aspect of assessment that has yet to mature in many aspects. We have made the first step in creating a procedural assessment of the laparoscopic cholecystectomy by connecting a list of procedural key steps of the laparoscopic cholecystectomy established with the Delphi method to a scale of independency. The independence-scaled assessment has a scale that connects to the control management used to guard patient safety by the supervising surgeon in the OR (Figure 2).

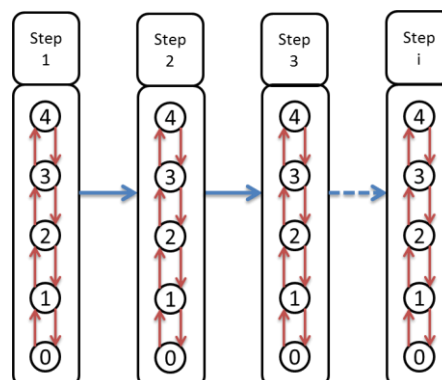


Figure 2: Interaction between intra-operative safety control dynamics, introduction of stepwise autonomy and procedural assessment. 0: Did not perform the step, 1: Able to perform a part of the task, 2: Performs the task with much guidance and instructions, 3: Performs the task with minimal guidance and instructions, 4: Can perform the whole task independent, safe and skilful. Red: within key step control management by supervising surgeon, blue: between key step control management by supervising surgeon.

The study results demonstrate a higher discriminative validity and inter-rater reliability of independence-scaled procedural assessment compared with the assessment with global rating scales. Our research group is currently conducting a study to define a cut-off score with an

acceptable sensitivity and specificity in identifying trainees who can be labelled as competent and receive a certification for the independent treatment of uncomplicated disease. This would complete the structured stepwise training and assessment system of laparoscopic procedures depicted in figure 3. Further research will also be necessary to evaluate whether the higher discriminative validity and inter-rater reliability of independence-scaled procedural assessment are also present in the assessment of advanced laparoscopic procedures.

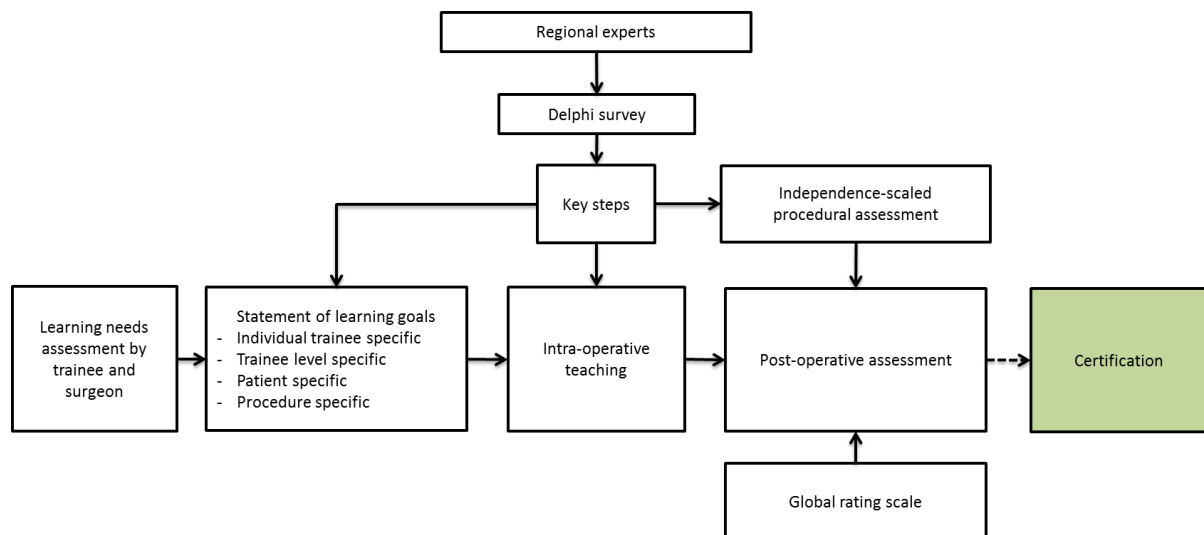


Figure 3: Structured curriculum for laparoscopic surgery training. Input of regional experts is used in a Delphi survey for the establishment of a list of key steps for procedural training and assessment for laparoscopic procedures. Delphi key steps are used for: 1) Pre-operative statement of learning goals, 2) Intra-operative teaching and a stepwise increase in autonomy 3) Post-operative assessment and 4) Certification for surgical treatment of uncomplicated disease.

After the publication of chapter 7 a sub-analyses was performed according to the guidelines in chapter 6, a chapter largely written after the publication of chapter 7 and 8. A correlation matrix was calculated to investigate whether the low reliability was somehow related to the (unrandomized) chronological order in which the video fragments were rated. The absolute agreement intra-class correlation coefficient (AA-ICC(2,1)) was calculated for video fragment 1 to 3 split into three groups: 1) early assessments, 2) middle assessments and 3) late assessments. In contrast to the earlier found low ICCs, the inter-rater reliability in the early groups were reasonable to good; 0.61 for fragment 1 ($p = 0.004$), 0.63 for fragment 2 ($p = 0.002$) and 0.42 for fragment 3 ($p = 0.016$) (Table 1). Also, the middle and late group ICCs were all non-significant and a consistent decrease was observed in the inter-rater reliability towards the late assessments.

Table 1: The AA-ICC's (2,1) of the early assessments, middle assessments and late assessments of the three video fragments F1-F3.

	Assessment		
	Early (N=54)	Middle (N=54)	Late (N=52)
F1 (N=51)	0.61*	0.27	0.19
F2 (N=53)	0.63*	0.35	0.03
F3 (N=56)	0.42*	0.24	0.04

* Statistical significant ICC-value ($p < 0.05$)

Raters involved in surgical education probably have a high drive to invest energy and time in the assessment of trainees. However, during the assessment, intrinsic motivation can be jeopardized by factors such as mental fatigue or time pressure. The results shown in table 1 indicate that these factors indeed may have decreased the accuracy in the assessment of surgical skills. In the subsequent evaluation of inter-rater reliability among assessors of surgical skills in chapter 8 we

therefore limited the assessments to an acceptable time frame and rewarded raters for the assessment of trainees to counteract any loss of motivation due to fatigue as much as possible. The above results should furthermore be seen as a warning to researchers and program directors who (unconsciously) overburden assessors during an attempt to gain surgical skills assessments.

Effect of training of assessors has been addressed in this thesis in chapters 6, 7 and 8. Teaching and assessing surgical raters is a field wherein much remains to be discovered. Questions like, what kind and how much training is necessary to make an assessment an accurate measurement of surgical skills, are still unanswered. An option would be to randomize a group of raters into two groups, a trained and untrained group, which uses a global rating scale and procedure-specific assessment to rate a series of performances.

Help from an experienced supervisor is crucial in the completion of a high risk complex task. It is the common perception that training with intensive training support leads to a higher performance level and faster attainment of proficiency during learning. Interestingly, psychologists have emphasized that learning should not be considered without taking in account the amount of retention of the learned skills. The degree of retention can differ significantly between learning methods. It has been shown in psychology that intensive guidance by trainers leads to higher performance level during training, but also to a higher decay of the acquired skills when the subject is to perform the same task in a later moment of time without the help of the trainer.^{5,6} This phenomenon has become known as the 'guidance hypothesis' and seems to be caused by the continuous provision of instructions. The constant verbalisation of the mind of the supervisor leads to insufficient free work capacity in the trainee to transform the work memory into chunks of information and store the chunks in long term memory.⁷ In our interaction with surgeons in our institution we have noticed that the prospect of an assessment that takes into account the amount of supervision a trainee actually needs to complete the operation induces a reticence in the supervising surgeon that would otherwise be absent. Restraining on the provision of cognitive support during a laparoscopic procedure might have the same diminishing effect on the decay of laparoscopic skills as has been observed in the retention of acquired complex motoric tasks in the field of educational psychology.⁵ Further research could provide information about whether this phenomenon also exists in surgical training.

References

- 1 Jalink MB, Goris J, Heineman E, Pierie J-PENP, ten Cate Hoedemaker HO. The effects of video games on laparoscopic simulator skills. *Am J Surg*. 2014;208(1):151–156.
- 2 Siska VB, Ann L, Gunter DW, Bart N, Willy L, Marlies S, *et al*. Surgical Skill: Trick or Trait? *J Surg Educ*. 2015;72(6):1247-1253.
- 3 Schumpeter JA. Vilfredo Pareto (1848-1923). *Q J Econ*. 1949;63:147–173.
- 4 Dijkstra FA, Bosker RJI, Veeger NJGM, van Det MJ, Pierie JPEN. Procedural key steps in laparoscopic colorectal surgery, consensus through Delphi methodology. *Surg Endosc*. 2014;29(9):1–8.
- 5 Schmidt R, Bjork R. New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychol Sci*. 1992;3(4):207-217.
- 6 Winstein CJ, Pohl PS, Lewthwaite R. Effects of physical guidance and knowledge of results on motor learning: support for the guidance hypothesis. *Res Q Exerc Sport*. 1994;65(4):316-323.
- 7 Van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: Design principles and strategies. *Med Educ*. 2010;44(1):85–93.

Chapter 10

Summary

Summary

The purpose of this thesis was to improve selection, training and assessment in laparoscopic surgery.

Assessment of candidates for medical specializations that require laparoscopic skills is subjective and lack scientific support. While some training institutions in dentistry, aviation and space exploration use aptitude test scores to obtain the optimal distribution of aptitude within their work force, studies on the predictive value of aptitude tests in laparoscopic surgery have been inconclusive about the value of aptitude assessment. Part I (Chapter 2) contains a review of the literature that describes the use of aptitude measurements to predict the acquisition and performance of laparoscopic skills. A meta-analysis was conducted to estimate the predictive power of 4 aptitude measurements: visual-spatial ability, psychomotor ability, perceptual ability and simulator-based assessment of aptitude. Although all aptitude tests showed a significant correlation (visual-spatial ability ($r = 0.32$; $p < 0.001$), perceptual ability ($r = 0.31$; $p < 0.001$) and psychomotor ability ($r = 0.26$; $p = 0.003$)), the highest correlation was observed for simulator-based assessment ($r = 0.64$; $p < 0.001$). Moreover, simulators are nowadays widely available in surgical departments involved in surgical training. The most straightforward option would therefore be to use these devices for assessment of aptitude for laparoscopic skills in a 'laparoscopy aptitude test'. It is important to keep in mind that medical knowledge, communication skills, decision-making skills and clinical judgment are core clinical competencies that should always be considered in conjunction with technical abilities when surgical competence is assessed in candidates.

Part II focuses on training laparoscopic skills in the OR. Training of laparoscopic skills has not been standardized and a large part of the learning process is completed in the high risk environment of the OR.

Chapter 3 describes the identification of common pitfalls during a laparoscopic cholecystectomy. In the ideal scenario, the whole learning curve for procedural learning would be completed in a simulator. However, current simulator computed measurements of improvement and proficiency criteria are often solely based on psychomotor skills such as time taken, instrument path length and number of collisions of the instruments with objects in the simulated work environment. These metrics do not reflect the full spectrum of vital elements for skilful and safe laparoscopic surgery in patients. We used the Pareto analysis to perform a training needs analysis of in vivo laparoscopy. We identified 11 aspects of trainee behaviour that account for 80% of verbal corrections given by supervising surgeons. These included behaviours like exercising the right amount of traction in the right direction with the (non-dominant) left-hand, choosing the right dissection plane and insertion of trocars on the right place and in the right direction. By conducting this analysis we have demonstrated that the Pareto-analysis can be seen as a highly potential method for calibrating laparoscopy skills lab training to on-the-job challenges. However, we have only demonstrated this for the laparoscopic cholecystectomy. Other procedures will also have to be analysed in order to establish content criteria for the whole scope of surgical procedures in laparoscopic surgery. Furthermore, Pareto-analysis based education methods should be evaluated to see whether this kind of training-needs-analysis leads to a higher training efficiency.

Chapter 4 compares two commonly used operation positions for the laparoscopic cholecystectomy, the French and the American position. The operation position used for training is the operation position that a trainee will probably use for the rest of his/her career. There is discussion over whether the French position has an ergonomic advantage in respect to the American position. Because of the difference in orientation of the surgeon towards the work field one would expect that the French position should be preferred. In our study of the surgeon's posture during the laparoscopic cholecystectomy no statistically significant difference was found between the French and American

position in terms of cervical spine flexion/extension ($p = 0.273$), thoracolumbar spine flexion/extension ($p = 0.273$), cervical spine torsion ($p = 0.715$), thoracolumbar spine torsion ($p = 0.465$), cervical spine lateroflexion ($p = 0.144$), or thoracolumbar spine lateroflexion ($p = 0.465$). No statistically significant difference was found in terms of the time spent within ergonomic acceptable angles in the sagittal plane for the cervical spine (French position, 71.5%; American position, 71.5%; $p = 0.273$) and the thoracolumbar spine (French position, 97.5%; American position, 95.1%; $p = 0.715$), the horizontal plane in the cervical spine (French position, 97.0%; American position, 82.8%; $p = 0.144$) and the thoracolumbar spine (French position, 94.7%; American position, 98.6%; $p = 0.144$) and the coronal plane in the cervical spine (French position, 98.4%; American position, 97.0%; $p = 0.715$) and the thoracolumbar spine (French position, 98.3%; American position, 97.4%; $p = 1.000$). Our results therefore indicate that, in the MIS suite, it does not seem to matter for the posture of the vertebral column whether the French or American position is used for the laparoscopic cholecystectomy. This is most likely a consequence of the presence of movable monitors in the MIS suite. However, more research is necessary to identify the aspects of the upper extremities which may especially be at risk of being overburdened in the American position in the MIS suite considering the position of the surgeon in this operation setup.

Chapter 5 states the key steps for 2 standardized laparoscopic procedures. This study uses a validated method, the Delphi method, to reach consensus among a group of twenty-one experts in laparoscopic surgery about which procedural steps should be seen as key steps for the laparoscopic cholecystectomy and appendectomy. Consensus was observed after the first round of Delphi on the key steps for laparoscopic appendectomy (Cronbach's alpha 0.92) and laparoscopic cholecystectomy (Cronbach's alpha 0.90). After the second round, 15 proposed key steps for laparoscopic appendectomy and 30 proposed key steps for laparoscopic cholecystectomy were rated as important (importance score $\geq 4/5$) by at least 80% of the expert panel. These key steps will be used in standardized training for trainees in the North-East surgical school and were used to create a procedure-based assessment for the laparoscopic cholecystectomy. The procedure-based assessment was evaluated in part III.

In Part III, we focus on the current subjectivity in the assessment of surgical skills and the absence of standardized methods for the assessment and certification for procedural skills.

Chapter 6 is focused on important aspects of the psychometrics behind inter-rater reliability. Twenty fold differences have been reported between the results of the 6 different mathematical models that can be used to calculate the intra-class correlation coefficient (ICC), a commonly used reliability coefficient for inter-rater reliability. This is a problem in current research about surgical education, as the majority of studies addressing the reliability of surgical skills assessment do not report which mathematical model was used. Second, there are some important issues pertinent in the evaluation of study quality in reliability research. Some of these have similarities with drug research, such as blinding of observers and random sequence generation, but manifest differently in the assessment of surgical skills. Third, the correct way of interpretation of the ICC is dependent on the purpose of the measurement instrument. Cut-off values, confidence intervals and probability distributions are all options that can be considered. Furthermore, it is important to take into account the constructivist social-psychological approach to assessment when interpreting surgical skills assessment. This means that different interpretation can sometimes be seen as equally valid as they are based on the individual professional experience, knowledge and socialization of the surgeon.

Chapter 7 describes the validity and reliability with a global rating scale (GRS) especially designed to assess laparoscopic skills, the Global Operative Assessment of Laparoscopic Skills (GOALS). GOALS was used to assess blinded randomized video fragments of a laparoscopic cholecystectomy. We used recordings of 6 consecutive cholecystectomies performed by 10 trainees. Out of the video recordings 3 fragments were edited to produce a total of 160 video fragments which were randomized and

assessed by two blinded laparoscopic surgeons previously unexposed to GOALS. Our study supports the existing evidence that GOALS has construct and concurrent validity for assessment of novice trainees performing a laparoscopic cholecystectomy. However, the reliability observed in this study was low (ICC=0.37) compared to the reliability found in other studies. There are a number of causes that could have been responsible for the low reliability: a lack of intrinsic/extrinsic motivation of the raters, fatigue of the raters, the lack of training in assessment and the characteristics of the Likert scale used in GRSs. The findings have led to a more rigorous methodological approach in the study of surgical assessment described in chapter 8.

Chapter 8 describes a new method for assessing procedural learning: independence-scaled procedural assessment. A procedural assessment system for a basic laparoscopic procedure has been developed by linking the key steps of the laparoscopic cholecystectomy to a scale of independency. The scale consists of 5 different levels for every step: 0) Did not perform the step, 1) Able to perform a part of the task, 2) Performs the task with much guidance and instructions, 3) Performs the task with minimal guidance and instructions, 4) Can perform the whole task independently, safely and skilfully. The procedural assessment was compared with 2 GRSs, the Objective Structured Assessment of Technical Skills (OSATS) and GOALS, in terms of validity, reliability and support among 10 surgeons and 6 scrub nurses. The participants rated blinded and subtitled full procedural videos of: 1) a novice trainee 2) an intermediate trainee and 3) a subcompetent trainee. Because of our findings in chapter 7, we attempted to calibrate the participants by showing them a short video of a laparoscopic performance of the low- and high-end of the scales of the GRSs. In contrast to the procedural assessment, the GRSs were not able to differentiate the intermediate from the subcompetent trainee. Thus, the discriminative validity of the procedural assessment was higher than for the GRSs. Furthermore, the surgeons showed a good reliability for the GRSs (OSATS 0.78; $p < 0.05$ and GOALS 0.74; $p < 0.05$), but an almost perfect reliability for the procedural assessment (0.84; $p < 0.05$). A survey that was distributed together with the surgical assessment forms showed that most surgeons were of opinion that the independence-scaled procedural assessment: 1) gives a more detailed picture of procedural skills than the GRSs, 2) is more objective than the GRSs and 3) should be reproduced for other laparoscopic procedures than the laparoscopic cholecystectomy. These findings indicate that the independence-scaled procedural assessment is a candidate that meets up to the requirement of an assessment tool for post-operative formative feedback, but perhaps also for high-stakes examinations such as certification. Furthermore, the reliability coefficients increased when the ratings of scrub nurses were added to those of the surgeons, indicating that scrub nurses can reliably assess the procedural laparoscopic skills of surgical trainees.

Interestingly, the step 'the dissection of Calot's triangle', displayed a moderate reliability (0.50). This key step has been identified as the most difficult subtask of the laparoscopic cholecystectomy and therefore demands a set of complex technical behaviours. In the light of constructivist social-psychological approach, raters will therefore focus on different aspects based on their knowledge, experience and previous socialization, causing a lower reliability than for the less difficult key steps. This also means that, although the ratings do not agree in the assessment of the dissection of Calot's triangle, they might all be equally valid, because they are based on the unique professional experience and understanding of the individual assessors. However, further quantitative, but definitively also qualitative research, is necessary to investigate whether this decrease in reliability can be thwarted by (practical) adjustments in the method of procedural assessment.

Samenvatting

Het doel van dit proefschrift is de verbetering van de selectie, training en beoordeling in de laparoscopische chirurgie.

De beoordelingswijze van kandidaten voor medische specialisaties die deels afhankelijk zijn van laparoscopische vaardigheden is subjectief en onvoldoende wetenschappelijk onderbouwd. Hoewel sommige educatieve instellingen in de tandheelkunde, luchtvaart en ruimtevaart al langere tijd minimumscores op neuropsychologische testen hanteren, zijn de uitkomsten van studies naar de voorspellende waarde van deze testen in de laparoscopische chirurgie niet eenduidig. Deel I (Hoofdstuk 2) van dit proefschrift bevat een overzicht van de literatuur die de voorspellende waarde van deze neuropsychologische testen heeft onderzocht in de laparoscopische chirurgie. Er werd een meta-analyse uitgevoerd om de voorspellende waarde te evalueren van 4 neuropsychologische testen: ruimtelijk-inzicht, perceptuele vaardigheden, psychomotorische vaardigheden en simulator vaardigheden. Hoewel al deze neuropsychologische testen een significante correlatie vertoonden met laparoscopische chirurgie (ruimtelijk-inzicht ($r = 0.32$; $p < 0.001$), perceptuele vaardigheden ($r = 0.31$; $p < 0.001$) en psychomotorische vaardigheden ($r = 0.26$; $p = 0.003$)), werd de hoogste correlatie gevonden tussen laparoscopische chirurgie en laparoscopische vaardigheden gemeten op een simulator ($r = 0.64$; $p < 0.001$). Simulators zijn tegenwoordig wijdverspreid beschikbaar op chirurgische afdelingen die betrokken zijn bij de educatie van laparoscopische chirurgie. Daarom is het gebruik van deze simulators de meest voor de hand liggende methode voor de beoordeling van de geschiktheid voor het leren van laparoscopische vaardigheden. In het kader van een sollicitatieprocedure voor een opleidingsplek tot chirurg moet de uitkomst van een neuropsychologische test of van laparoscopische vaardigheden op een simulator natuurlijk altijd worden beoordeeld in samenspraak met andere competenties (medische kennis, communicatieve vaardigheden, beslisvaardigheden, etc.).

Deel II is gefocust op de training van artsen in opleiding tot chirurg op de operatiekamer. De training op het gebied van laparoscopische vaardigheden is op dit moment niet gestandaardiseerd en een groot deel van het leerproces vindt plaats in de risicovolle leeromgeving van een operatiekamer.

Hoofdstuk 3 geeft een beschrijving van de verbale correcties gegeven door supervisors tijdens een laparoscopische cholecystectomie. In het ideale geval zou een arts in opleiding tot chirurg de gehele leercurve voor procedurele training doorlopen in een simulator. Echter, de huidige prestatiemetingen op simulators zijn vaak enkel gebaseerd op psychomotorische vaardigheden zoals de gebruikte tijd voor het voltooien van een taak, de afgelegde afstand van de instrumenten en het aantal botsingen tussen de instrumenten. Deze meetwaarden weerspiegelen slechts een klein deel van het spectrum aan vaardigheden dat nodig is voor veilige laparoscopische chirurgie op de operatiekamer. De resultaten van onze analyse toonden aan dat 11 aspecten van het gedrag van artsen in opleiding tot chirurg op de operatiekamer verantwoordelijk zijn voor 80% van het aantal verbale correcties gegeven door supervisors. Onder deze 11 aspecten vallen onder andere het uitvoeren van tractie met de (niet-dominante) linkerhand in de juiste richting en met de juiste kracht, het kiezen van een correct snijvlak en het inbrengen van de trocars op de juiste plaats en in de juiste richting. We demonstreerden middels deze analyse dat de Pareto-analyse kan worden gezien als een methode om laparoscopie training in het skillslab beter te kalibreren op de uitdagingen die artsen in opleiding tot chirurg tegenkomen in de operatiekamer. In de toekomst moeten op het Pareto-principe gebaseerde onderwijsmethoden geëvalueerd worden om te bepalen of deze analyse van trainingsbehoefte werkelijk leidt tot een efficiëntere training.

In hoofdstuk 4 worden twee vaak gebruikte opstellingen voor het uitvoeren van een laparoscopische cholecystectomie, 'the French position' en 'the American position', met elkaar vergeleken. De

operatiepositie die wordt aangeleerd tijdens de training is vaak de operatie positie die een arts in opleiding tot chirurg de rest van zijn of haar leven zal gebruiken. Vanwege het verschil in oriëntatie van de chirurg t.o.v. het werkveld zou men verwachten dat the French position beter is dan the American position. Bij vergelijking van de wervelkolom van de chirurg in de twee operatieposities werd er geen significant verschil gevonden in cervicale flexie/extensie ($p = 0.273$), thoracolumbaire flexie/extensie ($p = 0.273$), cervicale torsie ($p = 0.715$), thoracolumbaire torsie ($p = 0.465$), cervicale lateroflexie ($p = 0.144$), of thoracolumbaire lateroflexie ($p = 0.465$). Er werd tevens geen significant verschil gevonden in de hoeveelheid tijd binnen een ergonomisch acceptabele houding in het sagittale vlak van de cervicale wervelkolom (French position, 71.5%; American position, 71.5%; $p = 0.273$) en de thoracolumbaire wervelkolom (French position, 97.5%; American position, 95.1%; $p = 0.715$), het horizontale vlak van de cervicale wervelkolom (French position, 97.0%; American position, 82.8%; $p = 0.144$) en de thoracolumbaire wervelkolom (French position, 94.7%; American position, 98.6%; $p = 0.144$) en het coronale vlak van de cervicale wervelkolom (French position, 98.4%; American position, 97.0%; $p = 0.715$) en de thoracolumbaire wervelkolom (French position, 98.3%; American position, 97.4%; $p = 1.000$). Deze resultaten wijzen erop dat, in een operatiekamer ingericht voor minimale invasieve chirurgie, het niet uitmaakt voor de wervelkolom of de operatie in the French position of the American position wordt uitgevoerd. Het meest waarschijnlijk is dit een gevolg van de verplaatsbare monitors in operatiekamers aangepast voor laparoscopische chirurgie. Hoewel the American position voor de wervelkolom even ergonomisch blijkt te zijn als the French position, kunnen we geen uitspraken doen over de houding van de ledematen. Gezien de positie van de chirurg in the American position kunnen juist de armen en schouders at risk zijn voor overbelasting in deze operatieopstelling. Meer onderzoek is nodig om te evalueren of the French position en the American position dezelfde mate van comfort bieden aan de bovenste extremiteiten tijdens een laparoscopische cholecystectomie.

In hoofdstuk 5 worden de key steps van twee procedures beschreven die standaard met laparoscopische technieken worden uitgevoerd. In deze studie wordt een gevalideerde methode, de Delphi-methode, gebruikt om consensus te bereiken tussen 21 experts over welke stappen van de laparoscopische cholecystectomie en appendectomie tot de key steps van de procedure behoren. Er werd overeenstemming bereikt over key steps in de eerste ronde voor de laparoscopische appendectomie (Cronbach's alpha 0.92) en de laparoscopische cholecystectomie (Cronbach's alpha 0.90). Na de tweede ronde werden er 15 key steps voor de laparoscopische appendectomie en 30 key steps voor de cholecystectomie beoordeeld als belangrijk (score $\geq 4/5$) door minimaal 80% van het expertpanel. Deze key steps zullen worden gebruikt voor gestandaardiseerde training en beoordeling voor artsen in opleiding tot chirurg in Noordoost-Nederland. Een procedure specifieke beoordeling gebaseerd op deze key steps werd geëvalueerd in deel III.

In deel III wordt de subjectiviteit in de huidige beoordeling van operatieve vaardigheden aangekaart en wordt er gezocht naar een praktische methode voor het evalueren en feedback geven op het gebied van procedure specifieke vaardigheden.

Hoofdstuk 6 beschrijft de belangrijke aspecten van de psychometrie achter het concept inter-beoordelaars betrouwbaarheid. Er zijn twintigvoudige verschillen gerapporteerd in de inter-beoordelaars betrouwbaarheid berekend met de 6 verschillende berekenmodellen van de intra-class correlation coefficient (ICC), een veelgebruikte coëfficiënt voor het berekenen van de inter-beoordelaars betrouwbaarheid. Dit lijkt een probleem te zijn in onderzoek naar chirurgische training gezien het feit dat in het merendeel van de gevallen niet wordt beschreven welk model gebruikt is voor het berekenen van de ICC. Ten tweede zijn er een aantal problemen in de evaluatie van de kwaliteit van betrouwbaarheidsonderzoek in chirurgische training. Sommige van deze kwaliteitsaspecten hebben overeenkomsten met onderzoek naar medicijnen, zoals het randomiseren en het blinderen van participanten, maar manifesteren zich anders in inter-beoordelaars betrouwbaarheidsonderzoek. Ten derde is de correcte interpretatie wijze van de ICC afhankelijk van

het doel van het beoordelingsinstrument. Cut-off waarden, betrouwbaarheidsintervallen en waarschijnlijkheidsdistributies zijn opties die kunnen worden overwogen bij de interpretatie. Het is daarnaast belangrijk om een constructivistisch sociaalpsychologisch perspectief in beschouwing te nemen. Vanuit dit perspectief wordt onder andere beargumenteerd dat verschillende beoordelingen van supervisoren beschouwd kunnen worden als gelijkwaardig, terwijl deze beoordelingen kwantitatief van elkaar verschillen. Dit kan omdat de beoordelingen zijn gebaseerd op de individuele professionele ervaring, kennis en socialisatie van de beoordelende chirurg.

Hoofdstuk 7 beschrijft de validiteit en betrouwbaarheid van een Global Rating Scale (GRS) die speciaal ontwikkeld is voor de beoordeling van laparoscopische vaardigheden, namelijk de Global Operative Assessment of Laparoscopic Skills (GOALS). GOALS werd gebruikt om geblindeerde gerandomiseerde videofragmenten van een laparoscopische cholecystectomie te beoordelen. De video-opnames van 6 opeenvolgende laparoscopische cholecystectomiën uitgevoerd door 10 artsen in opleiding tot chirurg werden gebruikt om telkens fragmenten van 3 delen van de operatie te creëren. In totaal konden 160 gerandomiseerde videofragmenten worden beoordeeld door 2 geblindeerde laparoscopische chirurgen die geen eerdere ervaring hadden met GOALS. Onze studie ondersteunt het huidige bewijs dat GOALS een valide beoordelingsmethode is voor de beoordeling van beginnende artsen in opleiding tot chirurg die een laparoscopische cholecystectomie uitvoeren onder supervisie. Echter, de betrouwbaarheid was in deze studie laag (ICC = 0.37) vergeleken met andere studies. Er zijn een aantal oorzaken die hiervoor verantwoordelijk zouden kunnen zijn: intrinsieke/extrinsieke motivatie van de beoordelaars, vermoeidheid, een gebrek aan training in het gebruik van GOALS en/of de Likert-schaal die gebruikt wordt in GRSs. Deze bevinding is in ieder geval een motivatie geweest voor het gebruik van een meer rigoureuze methodologie in de studie naar de beoordeling van chirurgische vaardigheden die is beschreven in hoofdstuk 8.

Hoofdstuk 8 beschrijft een nieuwe methode voor de beoordeling van procedurele vaardigheden: de onafhankelijkheid geschaalde procedurele beoordeling. In deze studie werd een procedurele beoordeling ontwikkeld door de met de Delphi methode opgestelde key steps te koppelen aan een schaal van onafhankelijkheid. De schaal bestaat uit vijf verschillende niveaus voor elke stap: 0) heeft de stap niet uitgevoerd, 1) is in staat een deel van de stap uit te voeren, 2) voert de stap uit met veel begeleiding en instructies, 3) voert de stap uit met minimale begeleiding en instructies en 4) kan de gehele taak onafhankelijk, veilig en vaardig uitvoeren. De procedurele beoordeling werd vergeleken met twee GRSs, de Objective Structured Assessment of Technical Skills (OSATS) en de GOALS, in termen van validiteit, betrouwbaarheid en support onder 10 chirurgen en 6 OK-assistenten. De participanten beoordeelden geblindeerde en ondertitelde video's van drie artsen in opleiding tot chirurg: 1) een beginner, 2) een gevorderde beginner en 3) een bijna competente persoon in het uitvoeren van de laparoscopische cholecystectomie. Met het oog op de bevindingen in hoofdstuk 7 probeerden we de participanten te kalibreren door ze een korte video te tonen van laparoscopische vaardigheden op het laagste en het hoogste niveau van de schalen van de GRSs. In tegenstelling tot de procedurele beoordeling, waren de GRSs niet in staat een onderscheid te maken tussen de gevorderde beginnende arts en de bijna competente arts in opleiding tot chirurg. Het discriminerend vermogen van de procedurele beoordeling was dus hoger dan die van de GRSs. Verder werd er een goede inter-beoordelaars betrouwbaarheid gevonden voor de GRSs (OSATS 0.78; $p < 0.05$ en GOALS 0.74; $p < 0.05$), maar de inter-beoordelaars betrouwbaarheid was bijna perfect voor de procedure specifieke beoordeling (0.84; $p < 0.05$). Een enquête gedistribueerd onder de participanten samen met de beoordelingsformulieren toonde aan dat de meeste chirurgen van mening waren dat de onafhankelijkheid geschaalde procedurele beoordeling: 1) een beter beeld geeft van procedurele vaardigheden dan de GRSs, 2) objectiever is dan de GRSs en 3) zou moeten worden gereproduceerd voor andere laparoscopische procedures dan de cholecystectomie. Deze bevindingen lijken erop te wijzen dat de onafhankelijkheid geschaalde procedurele beoordeling een instrument is dat voldoet aan de eisen voor postoperatieve feedback, maar mogelijk ook voor zogenaamde 'high-stakes' beoordelingen, zoals certificatie. Ten slotte, werd de betrouwbaarheidscoëfficiënt hoger wanneer de

beoordelingen van de OK-assistenten werden toegevoegd aan die van de chirurgen. Dit wijst erop dat OK-assistenten een betrouwbare bron van feedback zijn op het gebied van operatieve vaardigheden en daarin een bijdragende rol kunnen spelen.

Het is interessant dat de stap 'dissectie van de driehoek van Calot' een matige betrouwbaarheid vertoonde (ICC = 0.50). Deze stap wordt gezien als de moeilijkste stap tijdens de procedure. Het voltooien van deze stap vereist op sommige momenten het simultaan uitvoeren van een set complexe technische vaardigheden. In het perspectief van de constructivistische sociaalpsychologische benadering richt iedere beoordelaar zich, op basis van hun eigen ervaring en socialisatie, daarom op verschillende aspecten van de vaardigheden die nodig zijn om deze stap veilig te voltooien. Waarschijnlijk leidt dit fenomeen tot een lagere inter-beoordelaars betrouwbaarheid dan voor de minder moeilijke stappen. Dus, hoewel de overeenstemming laag is, betekent dit niet per se dat de ene beoordeling beter is dan de ander. De beoordelingen kunnen alle een valide beeld geven van de vaardigheden die nodig zijn tijdens de dissectie van de driehoek van Calot, omdat ze alle gebaseerd zijn op unieke professionele ervaring. Verder kwantitatief, maar zeker ook kwalitatief, onderzoek is nodig om vast te stellen of deze verlaging in inter-beoordelaars betrouwbaarheid kan worden voorkomen door (praktische) aanpassingen in de wijze van beoordeling.

Appendix

Appendix A

The following formulas were used to convert correlation coefficients:

- In the case of subjective metrics, e.g. blinded assessment by experts, the reported inter rater reliability was used as an attenuation factor to calculate the corrected correlation (r_c) with the formula¹:

$$r_c = \frac{r_{obs}}{\sqrt{r_{apt}r_{sm}}} \quad (1)$$

where, r_{apt} = reliability of aptitude test (assumed to be equal to 1) and r_{sm} = reported reliability of subjective metric.

- In the case of non-parametric independent group comparisons with p-values < 0.05, the critical values of the Mann-Whitney U values were converted to r point-biserial (r_{pb}) with the formula²:

$$r_{pb} = 1 - \frac{2U}{n_1 n_2} \quad (2)$$

where, U = Mann Whitney U value for $\alpha_2 = 0.05$ and n_1 and n_2 are the number of participants in group 1 and 2. Because r_{pb} is a poor estimate of r, r_{pb} was consequently converted into r biserial (r_b) with the formula:

$$r_b = \frac{r_{pb} \sqrt{p_1 p_2}}{y} \quad (3)$$

where, p_1 and p_2 are the portions of group 1 and 2 and y is the y-value of the normal distribution at the z-score of the larger portion.

- When correlations were calculated with the Kendall tau rank correlation, the correlations were converted to the corresponding Pearson correlation with the formula³:

$$r_\tau = \sin(0.5\pi\tau) \quad (4)$$

- The critical Pearson correlation was calculated with the formula²:

$$r_{cv} = \sqrt{\frac{1}{\left(\left(\frac{N-2}{t^2}\right) + 1\right)}}$$

(5)

where, $t = t_{\text{critical value}}$ at $\alpha_2 = 0.05$ and $df = N - 2$.

To compensate for the partial independence between correlation because of the commonalities in study setting in which the different correlations were measured within a participant group, the following formula was used in the calculations of the mean variance (v_{group}) of the participant group effect sizes⁴:

$$V_{\text{group}} = \frac{v(1 + (m-1)r_x)}{m}$$

(6)

where, r_x is the correlation between the correlations of a participant group, m =number of correlations calculated with a particular group within a study and v =variance of the participant group effect size calculated on the basis of the number of participants within the group. If r_x is defined as 0, it means that the reported correlations can be seen as multiple independent studies. This leads to a decrease in the variance and can lead to overestimation of the precision of the summary correlation. A r_x value of 1 means that the correlations are entirely interdependent and leads to an underestimation of precision of the summary correlation. As no correlations could be identified in the literature that could be used to correct the partial interdependence between the study correlations, $r_x=0.5$ was used as a compromise between the two extremes.

1. Kock, A & Gemünden HG. A Guideline to meta-analysis. Retrieved April 1 from <https://www.tim.tu-berlin.de>.
2. De Coster J. Meta-analysis notes. Retrieved April 1 2015 from <http://www.stat-help.com/notes.html>.
3. Walker DA. JMASM9: converting Kendall's tau for correlational or meta-analytic analyses. *Journal of Modern Applied Statistical*. 2003;2(2);525-530.
4. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to meta-analysis*. John Wiley & Sons, Ltd.; West Sussex (England): 2009.

Appendix B

Author	Year	Reason for exclusion
Buckley	2013	Used a composite score of visual-spatial ability, perceptual ability and psychomotor ability
Buckley	2014	Used a composite score of visual-spatial ability, perceptual ability and psychomotor ability
Bartenbach	2014	Used only linear regression
Rosenthal	2006	Used only ANOVA to evaluate differences in learning plateau
Bruwaene	2014	Visual-spatial ability was only used to ascertain comparability between groups
Cadeddu	2003	Used non-linear causal resource analysis.
Utesch	2014	Technical difficulties with the simulator and/or misunderstanding of aptitude tests.
Hilgerink	2014	

Appendix C

Quality assessment based on QUADAS-2. ¹

Risk of bias		
Participant selection	Could the selection of participants have introduced bias?	1) Homogeneous group of participants (non-medical students/medical students/untrained trainees/trained trainees and consultants).
		2) >40 participants included or not significant correlations reported. ²
Index test	Could the conduct or interpretation of the index test have introduced bias?	Was the calculation method for the final score of the aptitude test not altered and was the calculation method reported?
Reference standard	Could the reference standard, its conduct, or its interpretation have introduced bias?	1) Was a validated method used to measure laparoscopic skills? (construct, concurrent or predictive validity of simulator metrics or subjective assessment has been shown within the study or previous literature)
		2) Was the performance score of laparoscopic skills not altered from the validated calculation method?
Flow and timing	Could the participant flow have introduced bias?	All recruited participants perform the aptitude test and laparoscopic skills measurement and were all included in the analysis or it is shown that the participants who completed their participation did not display different characteristics than the original group of participants.
Applicability		
Participant selection	Are there concerns that the included participants and setting do not match the review question?	Participants have interest in surgery or were motivated for participation by incentives. ³
Index test	Are there concerns that the index test, its conduct, or interpretation will not be applicable to the review question?	1) Was the execution of the aptitude test described in sufficient detail to permit replication of the test?
		2) Did the study describe whether response time was limited or not. ⁴
Reference standard	Are there concerns that the target condition as defined by the reference standard will not be applicable to the review question?	1) Was a validated method used to measure laparoscopic skills? (construct, concurrent or predictive validity of simulator metrics or subjective assessment has been shown within the study or previous literature)

Y=Yes, N=No, U=Unclear, -=not applicable, a=abstract.

Study			Risk of bias						Applicability			
Nr	Author	Year	Participant selection		Index test	Reference standard		Flow and timing	Participant selection	Index test		Reference standard
			1	2		1	2			1	2	
Visual-spatial ability												
1	Risucci ⁶	2000	N	Y	Y	Y	N	Y	Y	Y	N	N
2	Eyal ⁷	2001	Y	Y	Y	Y	Y	Y	Y	Y	N	N
3	Risucci ⁸	2001	N	Y	Y	Y	N	N	Y	Y	N	N
4	Haluck ⁹	2001	Y	Y	N	U	N	U	U	N	N	N
5a	Keehner ¹⁰	2004	N	Y	Y	Y	Y	U	Y	Y	Y	Y
5b	Keehner ¹⁰	2004	N	Y	Y	Y	Y	U	Y	Y	Y	Y
6	Schijven ¹¹	2004	Y	Y	Y	Y	Y	N	N	Y	Y	N
7	McClusky ¹²	2005	Y	Y	Y	Y	Y	N	Y	Y	Y	N
8	Stefanidis ¹³	2006	Y	Y	Y	Y	Y	N	Y	Y	N	N
9a	Hedman ¹⁴	2006	Y	Y	Y	Y	Y	Y	N	Y	Y	N
9b	Hedman ¹⁴	2006	Y	Y	Y	Y	Y	N	N	Y	Y	N

10	Keehner ¹⁵	2006	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
11	Birbas ¹⁶	2006	-	-	-	-	-	-	-	-	-	-
12	Andalib ¹⁷	2006	-	-	-	-	-	-	-	-	-	-
13	Hassan ¹⁸	2007	U	U	Y	Y	Y	Y	N	Y	Y	N
14	Enochsson ¹⁹	2008	-	-	-	-	-	-	-	-	-	-
15	Rosenthal ²⁰	2010	N	Y	Y	N	N	U	Y	Y	N	N
16	Sliwinski ²¹	2010	U	N	N	Y	U	U	Y	N	N	N
17	Kolozsvari ³	2011	N	Y	Y	Y	Y	Y	N	Y	Y	N
18	Jungmann ²²	2011	Y	Y	Y	Y	Y	Y	N	Y	Y	N
19	Ahlborg ²³	2011	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
20	Schlickum ²⁴	2011	Y	Y	Y	Y	Y	U	N	Y	Y	N
21	Luursema ²⁵	2012	Y	N	Y	Y	U	N	N	Y	N	N
22a	Ahlborg ²⁶	2012a	Y	Y	Y	Y	Y	Y	Y	Y	Y	N
22b	Ahlborg ²⁶	2012b	Y	N	Y	Y	Y	Y	Y	Y	Y	N
23a	Nugent ²⁷	2012a	Y	Y	Y	Y	Y	Y	N	Y	Y	N
23b	Nugent ²⁷	2012b	Y	N	Y	Y	Y	Y	Y	Y	Y	N
23c	Nugent ²⁷	2012c	Y	N	Y	Y	Y	Y	Y	Y	Y	N
23d	Nugent ²⁷	2012d	Y	N	Y	Y	Y	Y	Y	Y	Y	N
23e	Nugent ²⁷	2012e	Y	N	Y	Y	Y	Y	N	Y	Y	N
23f	Nugent ²⁷	2012f	Y	Y	Y	Y	N	Y	N	N	Y	N
24	Nugent ²⁸	2012	Y	N	Y	Y	Y	Y	Y	N	Y	N
25a	Ahlborg ²⁹	2013	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
25b	Ahlborg ²⁹	2013	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
25c	Ahlborg ²⁹	2013	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
26	Groenier ³⁰	2014	9,13,14,21, 22,25,26,33 -35,37	Y	Y	Y	Y	N	N	Y	N	N
Perceptual ability												
1	Haluck ⁹	2002	Y	Y	-	U	U	U	U	-	-	Y
2a	Gallagher ³¹	2003	Y	Y	-	N	N	Y	Y	-	-	N
2b	Gallagher ³¹	2003	Y	Y	-	N	N	Y	Y	-	-	N
2c	Gallagher ³¹	2003	N	Y	-	N	N	Y	Y	-	-	N
2d	Gallagher ³¹	2003	Y	Y	-	N	N	Y	Y	-	-	N
3	McClusky ¹²	2005	Y	Y	-	Y	Y	N	Y	-	-	N
4	Stefanidis ¹³	2006	Y	Y	-	Y	Y	N	Y	-	-	N
5	Kolozsvari ³	2011	N	Y	-	Y	Y	Y	N	-	-	N
6a	Nugent ²⁷	2012a	Y	Y	-	Y	Y	Y	N	-	-	N
6b	Nugent ²⁷	2012b	Y	N	-	Y	Y	Y	Y	-	-	N
6c	Nugent ²⁷	2012c	Y	N	-	Y	Y	Y	Y	-	-	N
6d	Nugent ²⁷	2012d	Y	N	-	Y	Y	Y	Y	-	-	N
6e	Nugent ²⁷	2012e	Y	N	-	Y	Y	Y	N	-	-	N
6f	Nugent ²⁷	2012f	Y	Y	-	Y	Y	Y	Y	-	-	N
Psychomotor ability												
1	Schijven ¹¹	2004	Y	Y	Y	Y	Y	N	N	Y	Y	N
2	Stefanidis ¹³	2006	Y	Y	Y	Y	Y	N	Y	Y	Y	N
3a	Nugent ²⁷	2012a	Y	Y	Y	Y	Y	Y	N	Y	Y	N
3b	Nugent ²⁷	2012b	Y	N	Y	Y	Y	Y	Y	Y	Y	N
3c	Nugent ²⁷	2012c	Y	N	Y	Y	Y	Y	Y	Y	Y	N
3d	Nugent ²⁷	2012d	Y	N	Y	Y	Y	Y	Y	Y	Y	N

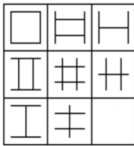

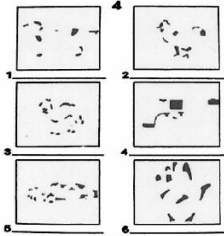


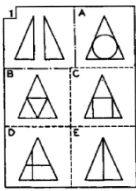
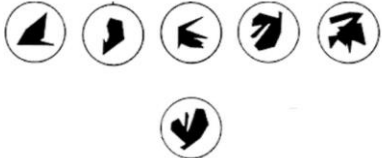
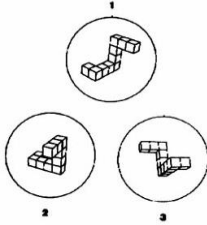
3e	Nugent ²⁷	2012e	Y	N	Y	Y	Y	Y	N	Y	Y	N
4	Nugent ²⁸	2012	Y	N	Y	Y	Y	Y	Y	N	Y	N
Simulator performance metrics												
1	Macmillan ³²	1999	Y	Y	Y	N	N	Y	Y	Y	-	Y
2a	Chaudhry ³³	1999	N	Y	Y	N	N	N	N	Y	-	N
2b	Chaudhry ³³	1999	N	Y	Y	N	N	N	Y	Y	-	N
2c	Chaudhry ³³	1999	Y	Y	Y	N	N	N	N	Y	-	N
3	Ahlberg ³⁴	2002	Y	Y	Y	Y	Y	N	Y	Y	-	Y
4	McClusky ¹²	2005	Y	Y	Y	Y	Y	N	Y	Y	-	N
5	Stefanidis ¹³	2006	Y	Y	Y	Y	Y	N	Y	Y	-	N
6	McCluney ³⁵	2007	N	Y	Y	Y	Y	U	Y	Y	-	Y
7	Hogle ³⁶	2008	Y	N	Y	Y	Y	Y	Y	Y	-	Y
8	Kundhal ³⁷	2009	Y	N	U	Y	Y	Y	Y	U	-	Y
9	Nugent ²⁷	2012	Y	N	Y	Y	Y	Y	Y	Y	-	N

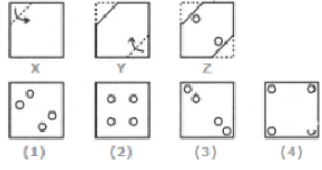
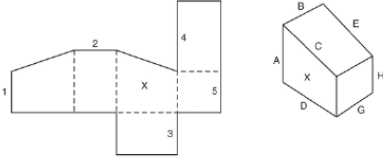
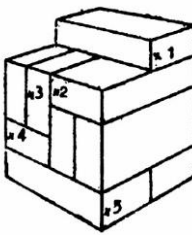
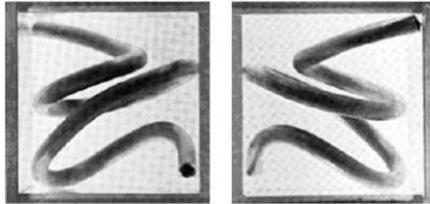
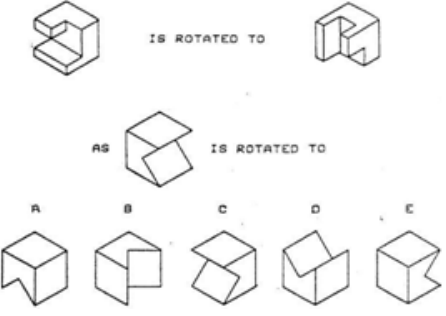
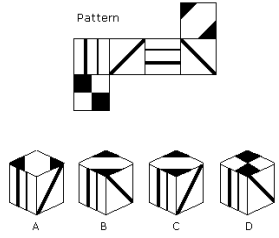
- Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011; 155:529–536.
- Wilson Van Voorhis CR, Morgan BL. Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology.* 2007;3:43-50.
- Kolozsvari N, Andalib A, Kaneva P, Cao J, Vassiliou M, Fried G, et al. Sex is not everything: the role of gender in early performance of a fundamental laparoscopic skill. *Surg Endosc.* 2011; 25:1037-1042.
- Partchev I, De Boeck P, Steyer R. How much power and speed is measured in this test? *Assessment.* 2013;20:242–52.
- Ritter FE, Baxter GD, Kim JW, Srinivasmurthy, S. Learning and retention. In: Lee JD & Kirlik A, editors. *The Oxford Handbook of Cognitive Engineering* (pp. 125-142). New York, NY: Oxford; 2013. P. 125-142.
- Risucci, Geiss, Gellman, Pinard, Rosser JC. Experience and visual perception in resident acquisition of laparoscopic skills. *Curr Surg.* 2000;57:368–372.
- Eyal R, Tendick F. Spatial ability and learning the use of an angled laparoscope in a virtual environment. *Stud Health Technol Inform.* 2001;81:146–152.
- Risucci D, Geiss A, Gellman L, Pinard B, Rosser J. Surgeon-specific factors in the acquisition of laparoscopic surgical skills. *Am J Surg.* 2001;181:289-93.
- Haluck RS, Gallagher AG, Satava RM. Reliability and validity of Endotower, a virtual reality trainer for angled endoscope navigation. *Stud Health Technol Inform.* 2002;85:179-84.
- Keehner MM, Tendick F, Meng MV, Anwar HP, Hegarty M, Stoller ML, et al. Spatial ability, experience, and skill in laparoscopic surgery. *Am J Surg.* 2004;188:71–75.
- Schijven MP, Jakimowicz JJ, Carter FJ. How to select aspirant laparoscopic surgical trainees: establishing concurrent validity comparing Xitact LS500 index performance scores with standardized psychomotor aptitude test battery scores. *J Surg Res.* 2004;121:112–119.
- McClusky DA, Ritter EM, Lederman AB, Gallagher AG, Smith CD. Correlation between perceptual, visuo-spatial, and psychomotor aptitude to duration of training required to reach performance goals on the MIST-VR surgical simulator. *Am Surg.* 2005;71:13-20.
- Stefanidis D, Korndorffer J, Black F, Dunne J, Sierra R, Touchard C, et al. Psychomotor testing predicts rate of skill acquisition for proficiency-based laparoscopic skills training. *Surgery.* 2006; 140:252-262.
- Hedman L, Ström P, Andersson P, Kjellin A, Wredmark T, Felländer-Tsai L. High-level visual-spatial ability for novices correlates with performance in a visual-spatial complex surgical simulator task. *Surg Endosc.* 2006;20:1275-1280.

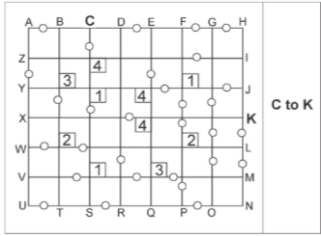
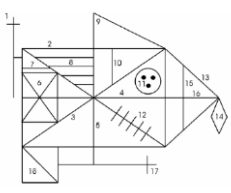
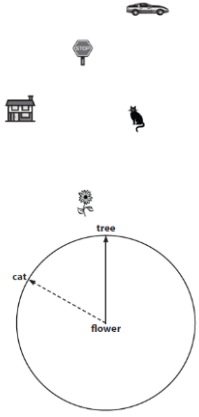
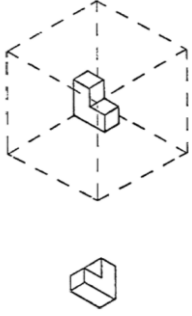
15. Keehner M, Lippa Y, Montello D, Tendick F, Hegarty M. Learning a spatial skill for surgery: how the contributions of abilities change with practice. *Appl Cognit Psychol*. 2006;20:487–503.
16. Birbas KN, Tzafestas CS, Kaklamanos IG, Vezakis AA, Polymeneas G, Bonatsos G. Spatial ability can predict laparoscopy skill performance of novice surgeons. 16th International Congress of the European Association for Endoscopic Surgery (EAES), Stockholm, Sweden, 11–14 June 2008.
17. Andalib A, Feldman LS, Cao J, McCluney AL, Fried GM. Can Innate visuospatial abilities predict the learning curve for acquisition of technical skills in laparoscopy? Scientific Session of the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), Dallas, Texas, USA, 26–29 April 2006.
18. Hassan I, Gerdes B, Koller M, Dick B, Hellwig D, Rothmund M, et al. Spatial perception predicts laparoscopic skills on virtual reality laparoscopy simulator. *Childs Nerv Syst*. 2007;23:685–689.
19. Enochsson L, Ahlborg L, Murkes D, Westman B, Hedman L, Kjellin A, Tsai-Felländer L. Visuospatial ability affects the performance of gynaecological simulation in the LapSimGyn® VR simulator. 16th International Congress of the European Association for Endoscopic Surgery (EAES) Stockholm, Sweden, 11–14 June 2008.
20. Rosenthal R, Gantert WA, Scheidegger D, Oertli D. Can skills assessment on a virtual reality trainer predict a surgical trainee’s talent in laparoscopic surgery? *Surg Endosc*. 2006;20:1286–1290.
21. Sliwinski J. Visuo-spatial ability and damage in laparoscopic simulator training. [Bsc thesis]. Tilburg, Noord-Brabant: Tilburg University; 2010; [cited 2015 Jan 7]. Available from: <http://essay.utwente.nl/60117/>.
22. Jungmann F, Gockel I, Hecht H, Kuhr K, Räsänen J, Sihvo E, Lang H. Impact of perceptual ability and mental imagery training on simulated laparoscopic knot-tying in surgical novices using a Nissen fundoplication model. *Scandinavian journal of surgery. Scand J Surg*. 2011; 100:78-85.
23. Ahlborg L, Hedman L, Murkes D, Westman B, Kjellin A, Felländer-Tsai L, Enochsson L. Visuospatial ability correlates with performance in simulated gynecological laparoscopy. *Eur J Obstet Gynecol Reprod Biol*. 2011;157:73-77.
24. Schlickum M, Hedman L, Enochsson L, Henningsohn L, Kjellin A, Felländer-Tsai L. Surgical simulation tasks challenge visual working memory and visual-spatial ability differently. *World J Surg*. 2011;35:710-715.
25. Luursema J-M, Verwey W, Burie R. Visuospatial ability factors and performance variables in laparoscopic simulator training. *Learning and individual differences*. 2012;22:632-638.
26. Ahlborg L, Hedman L, Rasmussen C, Felländer-Tsai L, Enochsson L. Non-technical factors influence laparoscopic simulator performance among OBGYN residents. *Gynecological Surgery*. 2012;9:415-420.
27. Nugent E. The evaluation of fundamental ability in acquiring minimally invasive surgical skill sets [MD thesis]. Dublin: Royal College of Surgeons in Ireland; 2012 [cited 2015 Jan 6]. Available from: <http://epubs.rcsi.ie/mdtheses/32/>.
28. Nugent E, Hseino H, Boyle E, Mehigan B, Ryan K, Traynor O, Neary P. Assessment of the role of aptitude in the acquisition of advanced laparoscopic surgical skill sets. *Int J Colorectal Dis*. 2012;27:1207-1214.
29. Ahlborg L, Hedman L, Nisell H, Felländer-Tsai L, Enochsson L. Simulator training and non-technical factors improve laparoscopic performance among OBGYN trainees. *Acta Obstet Gynecol Scand*. 2013;92:1194-201.
30. Groenier M, Schraagen J, Miedema H, Broeders I. The role of cognitive abilities in laparoscopic simulator training. *Advances in Health Sciences Education. Adv Health Sci Educ Theory Pract*. 2014;19:203-17.
31. Gallagher A, Cowie R, Jordan-Black J, Satava R, Crothers I. PicSOR: An objective test of perceptual skill that predicts laparoscopic technical skill in three initial studies of laparoscopic performance. *Surg Endosc*. 2003;17:1468-1471.

32. Macmillan AI & Cuschieri A. Assessment of innate ability and skills for endoscopic manipulations by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. *Am J Surg.* 1999;177:274–277.
33. Chaudhry AI, Sutton C, Wood J, Stone R, McCloy R. Learning rate for laparoscopic surgical skills on MIST VR, a virtual reality simulator: quality of human-computer interface. *Ann R Coll Surg Engl.* 1999;81:281-286.
34. 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;16:126–129.
35. McCluney AL, Vassiliou MC, Kaneva PA, et al. FLS simulator performance predicts intraoperative laparoscopic skill. *Surg Endosc.* 2007;21:1991-1995.
36. Hogle NJ, Widmann WD, Ude AO, Hardy MA, Fowler DL. Does training novices to criteria and does rapid acquisition of skills on laparoscopic simulators have predictive validity or are we just playing video games? *J Surg Educ.* 2008;65:431–435.
37. Kundhal PS & Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surg Endosc.* 2009;23: 645–649.

Appendix D

Intrinsic static																						
Matrix reasoning	<p>The participant is instructed to draw the next picture in the square right below:^{1,2}</p> 	Hidden patterns <p>The participant is instructed to identify the figures that contain the same shape as the figure to the left:³</p> 																				
Form completion	<p>The participant is instructed to identify the object displayed in the picture:³</p> 	Identical pictures <p>The participant is instructed to identify the object that's identical to the object to the left:³</p> 																				
Number comparison	<p>The participant is instructed to identify whether the number on the left is identical to those on the right:³</p> <table border="0" style="width: 100%;"> <tr> <td>659 ___ 659</td> <td>7343801 ___ 7343801</td> </tr> <tr> <td>73845 X 73855</td> <td>18824 ___ 18824</td> </tr> <tr> <td>1624 ___ 1624</td> <td>709216851 ___ 795816851</td> </tr> <tr> <td>438 X 436</td> <td>971 ___ 971</td> </tr> <tr> <td>4821459 ___ 4814259</td> <td>446014721 ___ 446014721</td> </tr> <tr> <td>698331 ___ 696331</td> <td>5173869 ___ 5172869</td> </tr> <tr> <td>11655 ___ 11658</td> <td>6430017 ___ 6430017</td> </tr> <tr> <td>617439428 ___ 617439428</td> <td>518198045 ___ 518168045</td> </tr> <tr> <td>1860439 ___ 1860439</td> <td>55179 ___ 55097</td> </tr> <tr> <td>90776105 ___ 90716105</td> <td>63216067 ___ 63216057</td> </tr> </table>		659 ___ 659	7343801 ___ 7343801	73845 X 73855	18824 ___ 18824	1624 ___ 1624	709216851 ___ 795816851	438 X 436	971 ___ 971	4821459 ___ 4814259	446014721 ___ 446014721	698331 ___ 696331	5173869 ___ 5172869	11655 ___ 11658	6430017 ___ 6430017	617439428 ___ 617439428	518198045 ___ 518168045	1860439 ___ 1860439	55179 ___ 55097	90776105 ___ 90716105	63216067 ___ 63216057
659 ___ 659	7343801 ___ 7343801																					
73845 X 73855	18824 ___ 18824																					
1624 ___ 1624	709216851 ___ 795816851																					
438 X 436	971 ___ 971																					
4821459 ___ 4814259	446014721 ___ 446014721																					
698331 ___ 696331	5173869 ___ 5172869																					
11655 ___ 11658	6430017 ___ 6430017																					
617439428 ___ 617439428	518198045 ___ 518168045																					
1860439 ___ 1860439	55179 ___ 55097																					
90776105 ___ 90716105	63216067 ___ 63216057																					
Intrinsic dynamic																						
Cards Rotation (2D)	<p>The participant is instructed to identify the identical objects on the right:³</p> 	Minnesota Paper Form Board (2D) <p>The participant is instructed to determine which complex design can be made by fitting together the simple geometric figures.⁴</p> 																				
Rotating shapes (2D)	<p>One of the upper shapes is rotated to make the shape below. The participant is instructed to identify whether it is reflected or not.⁵</p> 	Orientation <p>The participant is instructed to identify which of the shapes are the same:⁶</p> 																				

<p>Paper folding</p>	<p>The participant is instructed to identify the pattern that will appear when a piece of paper is folded and perforated in x, y and z:³</p> 	<p>Surface development</p>	<p>The participant is instructed to indicate which numbered site on the left shape corresponds with which letter on the right shape:³</p> 
<p>Block Touch</p>	<p>The participant is instructed to count the number of blocks touching each of the individual blocks:⁷</p> 	<p>Stumpf-Fay Cube Perspectives</p>	<p>Different views of 21 complex tubular figures have to be judged by the participant with respect to a specific point of view.⁸</p> 
<p>MRT-A (vertical rotation) MRT-C (vertical and horizontal rotation).</p>	<p>Same concept as orientation test. MRT-A contains rotations in vertical direction. MRT-C contains rotations in vertical and horizontal direction and is perceived as more difficult than MRT-A.⁶</p>	<p>Purdue Spatial Visualisation</p>	<p>The participant is instructed to identify which of the rotations are the same.⁹</p> 
<p>Cube Comparison</p>	<p>The participant is instructed to identify the cube that can be made with the object:³</p> 		

Extrinsic static		
Map Planning	<p>The participant is instructed to report the numbered box on the shortest line between the indicated letters. The shortest path cannot include a circle.³</p> 	<p>Rey figure</p> <p>The participant is instructed to copy the figure. After that the figure is drawn from memory immediately after withdrawal of the figure and minutes later.¹⁰</p> 
Extrinsic dynamic		
Perspective-taking	<p>The participant is instructed to imagine standing at the flower and facing the tree and to point to the cat.¹¹</p> 	<p>Visualization of views</p> <p>The participant is instructed to indicate which of the corners of the figure below is touching the corner of the cube drawn up.⁹</p> 

1. Raven JC. Progressive matrices: A perceptual test of intelligence. London: H. K. Lewis & Co; 1938.
2. Raven JC. Advanced Progressive Matrices. Sets I and II. London: H. K. Lewis & Co. 1965.
3. Ekstrom RB, French JW, Harman HH, Dermen D. Manual for kit of factor referenced cognitive tests. Princeton: Educational Testing Service; 1976.
4. Likert R, Quasha WH. Revised Minnesota Paper Form Board Test Manual. 2nd ed. San Antonio: Psychological Corporation; 1995.
5. Cooper LA. Mental rotation of random two-dimensional shapes. *Cogn Psychol*. 1975;7:20–43.
6. Vandenberg SG, Kuse AR. Mental rotations, a group test of three-dimensional spatial visualization. *Percept Mot Skills*. 1978;47:599–604.
7. Thurstone LL. Psychological tests for the study of mental abilities. Chicago: University of Chicago Press; 1937.
8. Stumpf H, Fay E. Schlauchfiguren. Ein Test zur Beurteilung des räumlichen Vorstellungsvermögens. Göttingen: Hogrefe; 1983.
9. Guay RB. Purdue spatial visualization test – visualization of rotations. West Lafayette: Purdue Research Foundation; 1976.
10. Osterrieth PA. Le test de copie d'une figure complexe: Contribution a l'étude de la perception et de la memoire [The Complex Figure Test: Contribution to the study of perception and memory]. *Arch Psychol*. 1944;28:1021–1034.
11. Kozhevnikov M, Hegarty M. A dissociation between object-manipulation and perspective-taking spatial abilities. *Mem Cognit*. 2001;29:745-756.

Appendix E

PMA test	Description	Outcome	Measure
Grooved Pegboard ^{1,2}	25 holes with randomly positioned slots that have a key along one side that must be filled with pegs.	Execution time, number of pegs dropped and number of pegs correctly placed in the holes for left and right hand.	<ol style="list-style-type: none"> 1. Gross movements of the fingers, hands, and arms. 2. Fine fingertip dexterity. 3. Left/right eye-hand coordination.
Purdue Pegboard ^{3,4}	25 holes that have to be filled as fast as possible with pins.	Number of pegs inserted within 30 seconds for left hand, right hand and for both hands.	<ol style="list-style-type: none"> 1. Gross movements of the fingers, hands, and arms. 2. Fine fingertip dexterity. 3. Bimanual coordination. 4. Left/right/bimanual coordination
Crawford Small Parts Dexterity ⁵	The participant tries to place pins into small holes in a plate with tweezers and fits collars over the pins. In the second part, the participant places small screws into threaded holes in the plate.	Execution time.	Motor control and eye-hand coordination.
Gibson Spiral Maze ⁶	The participant traces a line through a printed paper maze in the least amount of time and has to avoid obstacles while stress-enhancing triggers are administered.	Execution time and error score.	Eye-hand coordination.
Finger Tap ⁷	Participant repetitively taps a lever as fast as possible with one hand in 5 periods of 10.	Average number of taps for each hand.	Motor speed and lateralized coordination.
Tremor ⁸	Participant tries to hold a laparoscopic grasper holding a needle that is attached to a shaker as steady as possible.	The number of oscillations.	Hand steadiness.
Reaction time ⁸	Participant presses a button and must try to press one of three other buttons as fast as possible after it has lit up.	Time delay in response.	Response speed.

1. Klove H. Clinical neuropsychology. *The medical Clinics of North America*. 1963;47:1647-1658.
2. Trites RL. Neuropsychological Test Manual. Ottawa: Royal Ottawa Hospital; 1977.
3. Tiffin J & Asher EJ. The Purdue pegboard: norms and studies of reliability and validity. *The Journal of Applied Psychology*. 1948; 32:234-247.
4. Tiffin J. Purdue Pegboard Examiner's. Manual. Rosemont: London House; 1968.
5. Crawford JE & Crawford DM. Crawford Small Parts Dexterity Test: manual. San Antonio: Psychological Corporation; 1956.
6. Gibson HD. The Gibson Spiral Maze test: retest data in relation to behavioural disturbance, personality and physical measures. *Br J Psychol*. 1964;55:219-225.
7. Halstead WC. Brain and intelligence: a quantitative study of the frontal lobes. Chicago: University of Chicago Press; 1947.
8. Stefanidis D, Korndorffer J, Black F, Dunne J, Sierra R, Touchard C, et al. Psychomotor testing predicts rate of skill acquisition for proficiency-based laparoscopic skills training. *Surgery*. 2006;140:252-62.

Appendix F

Cluster	Verbale correction	Repeats per procedure	Step 1	Step 2-3	Step 4	Step 5	Step 6	Cumulative number	Cumulative percentage
1	Tensioning the gallbladder with the appropriate direction and strength	6.89		+	+	+		441	27
2	Identifying the correct surgical plane	4.75		+		+		745	46
3	Use of the dissection hook	1.66		+		+		851	52
4	Choosing position and direction of trocar placement	1.58	+					952	58
5	Using the clamp	1.05		+		+		1019	62
6	Staying close to the gallbladder	0.95		+		+		1080	66
7	Staying superficial during dissection	0.77		+		+		1129	69
8	Using the clipping instrument	0.73			+			1176	72
9	Avoiding harm to surrounding structures other than the liver	0.70		+		+		1221	75
10	Avoiding liver damage	0.66		+		+		1263	77
11	Positioning of the clip	0.61			+			1302	80
12	Use of the endobag	0.42					+	1329	81
13	Dissection towards a direction away from the gallbladder	0.39		+		+		1354	83
14	Hemostasis	0.38		+		+		1378	84
15	Use of the scissors	0.31		+	+			1398	86
16	Depth and width of incision	0.31	+					1418	87
17	Position for the start of dissection of the peritoneum/adhesiolysis	0.27		+				1435	88
18	Use of the crocodile clamp	0.27					+	1452	89
19	Instrument change	0.23		+	+	+		1467	90
20	Anatomy during dissection*	0.22		+				1481	91
21	Positioning of the patient in anti-trendelenburg	0.22	+					1495	92
22	Use of the suction instrument	0.22		+		+		1509	92
23	Preventing intra-abdominal injury during trocar placement	0.20	+					1522	93
24	Removal of the gallbladder	0.20					+	1535	94
25	Use of the foot paddle	0.19		+	+	+		1547	95
26	Anatomy during diagnostic laparoscopy*	0.19	+					1559	95
27	Removing the gallbladder out of sight	0.17				+		1570	96
28	Safe usage of the cautery	0.17		+				1581	97
29	Adhesiolysis	0.16		+				1591	97
30	Positioning the gallbladder	0.14					+	1600	98
31	Searching for an alternative approach when there is a stagnation in the progression	0.13		+		+		1608	98
32	Tissue handling	0.11				+		1615	99
33	Coordination of the instruments	0.09		+		+	+	1621	99
34	Localizing the gallbladder	0.05	+					1624	99
35	Technical aspects of creating pneumoperitoneum *	0.03	+					1626	100

36	Removals of trocars under direct sight	0.03					+	1628	100
37	Removal of stones	0.03					+	1630	100
38	Communication about technical questions about instruments	0.02					+	1631	100
39	Preventing perforation of the gallbladder	0.02					+	1632	100
40	Irrigation	0.02					+	1633	100

+ = behaviour addressed in procedural step * = theme of questions of supervisor about procedural knowledge

Dankwoord

Dit proefschrift is mede te danken aan de volgende mensen:

Prof. J.P.E.N. Pierie, Bedankt voor de kans die je mij gaf na het sturen van mijn e-mail. Je hebt de deur voor mij geopend naar een promotietraject. Op goede momenten heb je mij bemoedigd om door te gaan en op minder goede momenten ben je in mij blijven geloven. Als dingen moesten worden geregeld stond je altijd klaar om de zaken in orde te brengen. Het is een voorrecht om onder jouw leiding onderzoek te mogen doen.

Beste **Marc**, de inzet tijdens jouw promotieonderzoek was de basis voor mijn promotieonderzoek. Zonder jouw dataverzameling en begeleiding was het nooit gelukt. In principe ben je meer dan één keer gepromoveerd door mij jouw data te laten analyseren en zo grey literature tot white literature te maken.

Beste **Martijn Bethlehem, Kevin Wevers, Ilona Pereboom, Frederieke Dijkstra** en **Mirjam Keijzer**. Jullie inbreng voor het onderzoek waardeer ik zeer. Door jullie feedback is er een beoordelingsmethode tot stand gebracht waar ik trots op ben.

Mede-auteurs, **Christiaan Hoff, Erik Totte, Henk ten-Cate Hoedemaker**, dank voor jullie bijdrage aan dit proefschrift.

OK-assistenten **Ingeborg Riedstra, Wiep Rienks, Linda van de Meulen, Jeroen Kindt, Hindrik Boonstra, Fronnie Kramer, Gerda Kootstra, Lotte van der Werff**, jullie hebben aangetoond dat ervaring als OK-assistent ook nuttig kan zijn voor de beoordeling van aiosen. Vooral jullie hoge betrouwbaarheid in de items 'gebruik van instrumenten' en 'gebruik van assistentie' van de OSATS vond ik opvallend. Jullie zijn onmisbaar op de operatiekamer, maar misschien zouden jullie ook een essentieel onderdeel moeten worden van de feedback naar aiosen.

Beste **Nic Veeger**, ik heb veel met je gelachen over de waanzin van het berekenen van een Pearson correlatie voor het vaststellen van de inter-beoordelaars betrouwbaarheid. Voor de meta-analyse was jij de enige met wie ik een feedback gesprek kon voeren over mijn werkvaardigheden. Je kritische blik op de statistische analyse heeft me er meerdere keren toe gedwongen weer de boeken in te duiken en heeft de kwaliteit van de analyses daardoor verhoogd.

Beste **Marcel Dunand**, tijdens mijn promotietraject overkwam jou een vervelende aandoening. Ik wil je hier nogmaals bedanken voor jouw woorden van licht in tijden van duisternis.

Beste **Jerry Mendeszoon**, jarenlang ben je mijn geestelijke vader geweest. Jouw wijsheid en enthousiasme zijn mij altijd bijgebleven. Ik heb door jou 'mijn huis leren bouwen op de rots'.

Beste **Stephan Jonker**, door jouw inzet heb ik mijn chemie practica uiteindelijk goed kunnen afsluiten. Je hebt me bemoedigd met je positieve feedback op mijn proton NMR spectroscopie analyse, maar vooral je gevoel voor morele rechtvaardigheid heeft mijn carrière beïnvloed.

Beste **Atiev**, je bent voor een bijzondere leraar voor mij geweest in het KTC. Je hebt me bemoedigd met je warmte, enthousiasme, humor en vrijgevigheid. Ik heb het boek wat je me gaf met grote interesse gelezen.

Beste **Michline**, je bent een heldin! Bedankt voor je bemoediging!

Beste **Sina** en **Jeremie**, we waren de 3 musketiers tijdens onze studieperiode. Jammer genoeg kozen we andere wegen na onze studie. Jullie hebben gezorgd voor een aantal van de mooiste momenten van mijn leven. Bedankt voor jullie kritische blik op mijn onderzoek en voor de illustraties in het promotieboekje!

Beste **Arman** en **Yaser**, na 3 keer te zijn uitgeloofd met een propvolle auto door een stormachtige regenbui verhuizen naar Leuven, door de sneeuw in -10°C gasthuisberg op fietsen om in de stilte van de bibliotheek te kunnen voorbereiden op het toelatingsexamen, 12 uur per dag tussen de nonnen zitten in de faculteit theologie om maar niet afgeleid te worden door andere medische studenten tijdens het voorbereiden op een tentamen celfysiologie... Enkele voorbeelden van onze inspanningen. Gelukkig kunnen we er nu de vruchten van plukken.

Beste **Piem** en **Nele**, jullie waren nooit beroerd om verre afstanden te reizen voor ontspanning in de wat minder drukke momenten op onze agenda. Bedankt daarvoor!

Beste **Joel**, we hebben samen risico's genomen, jij voor je dochter en ik voor mijn studie. Voor mij was dat niet zonder negatieve gevolgen, maar ik ben blij dat ik er een goede vriend aan heb overgehouden.

Beste **Jean Paul, Jacquinot, Paul** en **Malcolm**, ik heb veel met jullie gelachen. Als broeders hebben jullie vooral Leeuwarden tot de leukste tijd van mijn promotieonderzoek gemaakt.

Beste **Maarten Jalink**, jouw proefschrift was het voorbeeld voor mijn proefschrift. Het was erg bijzonder om na het lezen van jouw proefschrift samen met jou als zaalarts te mogen werken. Ik heb daardoor academische, maar ook klinische vaardigheden van je geleerd! Bedankt voor je feedback op een aantal artikelen in dit proefschrift!

Beste **Armand van Kanten, Soeradj Harkisoen, Anuska Jewbali, Ted Nannan Panday, Rohiet Girjasing** en **arts-assistenten chirurgie**. Ik wens jullie veel succes met de medische zorg in Suriname en hoop in de toekomst nog wat voor het AZP te kunnen betekenen. Tang bun!

Beste **Ebi** en **Roos**, bedankt voor het nakijken van mijn resp. Engelse en Nederlandse taalfouten.

Beste **Fimke Heslinga, Anniek Boer, Maaïke Hetteema, Kor Hutting** en **Thijs Wind** jullie hebben hard gewerkt op de computers naast mij. Door mij als vogel in een vogelzwerm te conformeren aan jullie gedrag heb ik meer uit mijn promotieproject gehaald.

Beste **mam**, ik hou van je.

Curriculum Vitae

Kelvin Kramp was born in Rotterdam, the Netherlands, on October 31th 1984.

After graduating high school (Atheneum, Capelle aan den IJssel), he completed a bachelor chemistry at the University of Utrecht with a minor in psychology. After not getting into medicine by lottery three times in a row, he moved to Leuven in Belgium to keep pursuing a career in medicine. During the first and second year of studying abroad he participated in the admission test for the fast-track medicine at the University of Groningen. In the second year he succeeded in becoming one of the few selected for this trajectory and moved to Groningen. He completed his internships at the Medical Centre Leeuwarden and for his final internship he travelled back to the University Hospital Leuven to work there in the department of surgery and emergency medicine.

During his 2nd master he started doing research headed by MD/PhD M.J. van Det and prof. J.P.E.N. Pierie. This led to his first publication, which formed the start of a PhD trajectory at the Medical Centre Leeuwarden after completion of his study.

After his PhD trajectory he worked one year in the Academic Hospital of Paramaribo in Suriname at the department of surgery as a resident. He is currently working in the emergency department in the St. Antonius Hospital in Sneek.

List of publications

This thesis

- 2016 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
The Pareto-analysis for establishing content criteria in surgical training
J Surg Educ. 2016;73:892-901
- 2016 Kramp KH, van Det MJ, Veeger NJGM, Hoff C, ten Cate Hoedemaker HO, Pierie JP
The predictive value of aptitude assessment in laparoscopic surgery: a meta-analysis
Med Educ. 2016;50:409-427
- 2015 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
Validity, reliability and support for implementation of independence-scaled procedural assessment in laparoscopic surgery
Surg Endosc. 2016;30:2288-2300
- 2015 Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJ, Pierie JP
Validity and Reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in Novice Trainees Performing a Laparoscopic Cholecystectomy
J Surg Educ. 2015;72:351-358
- 2014 Kramp KH, van Det MJ, Totte ER, Hoff C, Pierie JP
Ergonomic assessment of the French and American position for laparoscopic cholecystectomy in the MIS Suite
Surg Endosc. 2014;28:1571-1578
- 2014 Bethlehem MS, Kramp KH, van Det MJ, ten Cate Hoedemaker HO, Veeger NJ, Pierie JP
Development of a standardized training course for laparoscopic procedures using Delphi methodology
J Surg Educ. 2014;71:810-816

Other publications

- 2016 Kramp KH, Spruit E, van Zwieten T, van Det MJ, Pierie JP
A systematic review of operating room teaching behavior
In progress
- 2016 Van Zwieten T, Kramp KH, van Det MJ, Pierie JP
The OSATS, GOALS and independence-scaled procedural assessment in teaching laparoscopic surgery
In progress
- 2016 Jalink M, Kramp KH, Baktawar S, Jewbali A
Skin necrosis after self-removal of a artificial penile nodule in a Surinamese man
BMJ Case Rep. 2016 Jun 28
- 2015 Kramp KH, Omer MG, Schoffski P, d'Hoore A

Sphincter sparing resection of a large obstructive distal rectal gastrointestinal stromal tumour after neoadjuvant therapy with imatinib (Glivec)
BMJ Case Rep. 2015 Jan 8

- 2015 Verbeek HH, Meijer JA, Zandee WT, Kramp KH, Sluiter WJ, Smit JW, Kievit J, Links TP, Plukker JT
Fewer Cancer Reoperations for Medullary Thyroid Cancer After Initial Surgery According to ATA Guidelines
Ann Surg Oncol. 2015;22:1207-1213

List of congress presentations

- 2016 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
The Pareto-analysis for establishing content criteria in surgical training.
16th World Congress of Endoscopic Surgery in Amsterdam, June 2016
Poster presentation
- 2016 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
The Pareto-analysis for establishing content criteria in surgical training.
2016 SAGES Congress in Boston, March 2016
Oral presentation
- 2015 Kramp KH, van Det MJ, Hoff C, Veeger NJGM, Ten Cate Hoedemaker HO, Pierie JP
A meta-analysis of aptitude measurement in laparoscopic surgery
15th World Congress of Endoscopic Surgery in Bucharest, June 2015
Oral presentation
- 2015 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
Validity, reliability and feasibility of the OSATS, GOALS and independency procedure-based
assessment in laparoscopic surgery
12de NVEC congress 'New Age Surgery', March 2015
Poster presentation
- 2015 Kramp KH, van Det MJ, Hoff C, Veeger NJGM, Ten Cate Hoedemaker HO, Pierie JP
A meta-analysis of aptitude measurement in laparoscopic surgery
12de NVEC congress 'New Age Surgery', March 2015
Poster presentation
- 2015 Kramp KH, van Det MJ, Hoff C, Veeger NJGM, Ten Cate Hoedemaker HO, Pierie JP
A meta-analysis of aptitude measurement in laparoscopic surgery
Medical Centre Leeuwarden scientific symposium, February 2015
Poster presentation
- 2015 Kramp KH, van Det MJ, Hoff C, Veeger NJGM, Ten Cate Hoedemaker HO, Pierie JP
A meta-analysis of aptitude measurement in laparoscopic surgery
Medical Centre Leeuwarden scientific symposium, February 2015
Poster presentation
- 2015 Kramp KH, van Det MJ, Veeger NJGM, Pierie JP
Validity, reliability and feasibility of the OSATS, GOALS and independency procedure-based
assessment in laparoscopic surgery
Medical Centre Leeuwarden scientific symposium, February 2015
Oral presentation
- 2014 Kramp KH, van Det MJ, Totte ER, Hoff C, Pierie JP
Ergonomic assessment of the French and American position for laparoscopic
cholecystectomy in the MIS Suite
Medical Centre Leeuwarden scientific symposium, February 2014
Poster presentation
- 2014 Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJGM, Pierie JP

Validity and reliability of Global Assessment of Laparoscopic Surgery (GOALS) in novice trainees performing a laparoscopic cholecystectomy
11de NVEC congres 'Hollandsche Meesters', March 2014
Poster presentation

2014 Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJGM, Pierie JP
Validity and reliability of Global Assessment of Laparoscopic Surgery (GOALS) in novice trainees performing a laparoscopic cholecystectomy
14th World Congress of Endoscopic Surgery in Paris, June 2014
Poster presentation

2014 Kramp KH, van Det MJ, Totte ER, Hoff C, Pierie JP
Ergonomic assessment of the French and American position for laparoscopic cholecystectomy in the MIS Suite
2014 SAGES Congress in Salt Lake City, april 2014
Poster presentation