Assessment of Laparoscopic Suturing Skills

Elif Bilgic BSc

Doctor of Philosophy

Department of Experimental Surgery

McGill University

Montreal, Quebec, Canada

January 2018

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy, Experimental Surgery

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Dr Melina C Vassiliou for her mentorship and all the opportunities she has provided me. Throughout the past years, I have gained great insight into this field and learnt from her extensive knowledge and experience. This thesis would not have been possible without her support at every stage.

I would like to thank Dr Yusuke Watanabe, whom I had the opportunity to work with side by side in developing my research ideas and skills. His collaboration and expertise has been key in my journey and evolution.

I am thankful to all the members of the Henry K.M. DeKuyper Education Centre, and the Steinberg- Bernstein Centre for Minimally-Invasive Surgery and Innovation, especially Drs Liane S Feldman and Gerald M Fried, for their endless support. I would also like to thank my friend Emily A Polak for helping me throughout the writing process of my thesis.

Last but not least, I am greatful to the encouragement and continuous love that I have received from my parents E Ayse Bilgic and Mehmet T Bilgic, and my sister K Merve Bilgic. This work would not have been possible without them.

Thank you so much everyone

CONTRIBUTION OF AUTHORS

I have made substantial contribution to all co-authored manuscripts in this thesis. Original research questions, and all of the stages of the studies were developed in collaboration with my thesis supervisor, Dr. Melina C Vassiliou. The individual contributions of the co-authors for each manuscript are described below (* = co-primary authors):

Bilgic E, Alyafi M, Landry T, Vassiliou MC. Simulation Platforms to Assess Laparoscopic Suturing: a Scoping Review. [Not yet submitted]

- *Study conception and design:* Bilgic, Vassiliou
- *Data acquisition:* Bilgic, Landry
- *Analysis and interpretation of data:* Bilgic, Alyafi, Vassiliou
- *Drafting of manuscript:* Bilgic, Vassiliou
- *Critical revision:* Bilgic, Alyafi, Vassiliou

Bilgic E, Kaneva P, Okrainec A, Ritter EM, Schwaitzberg SD, Vassiliou MC. Trends in the Fundamentals of Laparoscopic Surgery® (FLS) certification exam over the past 9 years. *Surgical Endoscopy*. 2017. [E pub ahead of print]

- *Study conception and design:* Bilgic, Okrainec, Ritter, Schwaitzberg, Vassiliou
- *Data acquisition:* N/A
- *Analysis and interpretation of data:* Bilgic, Kaneva, Okrainec, Ritter, Schwaitzberg, Vassiliou
- *Drafting of manuscript:* Bilgic, Kaneva, Vassiliou
- *Critical revision:* Bilgic, Kaneva, Okrainec, Ritter, Schwaitzberg, Vassiliou

Bilgic E, Watanabe Y*, Nepomnayshy D, Gardner A, Fitzgibbons S, Ghaderi I, Alseidi A, Stefanidis D, Paige J, Seymour N, McKendy KM, Birkett R, Whitledge J, Kane E, Anton NE, Vassiliou MC. Multicenter Proficiency Benchmarks for Advanced Laparoscopic Suturing Tasks. *American Journal of Surgery.* 2017;213:217-221.

- *Study conception and design:* Bilgic, Watanabe, Nepomnashy, Vassiliou
- *Data acquisition:* Bilgic, Watanabe, Nepomnayshy, Gardner, Fitzgibbons, Ghaderi, Alseidi, Stefanidis, Paige, Seymour, McKendy, Birkett, Whitledge, Kane, Anton, Vassiliou
- *Analysis and interpretation of data:* Bilgic, Watanabe
- *Drafting of manuscript:* Bilgic, Watanabe, Vassiliou
- *Critical revision:* Bilgic, Watanabe, Nepomnayshy, Gardner, Fitzgibbons, Ghaderi, Alseidi, Stefanidis, Paige, Seymour, McKendy, Birkett, Whitledge, Kane, Anton, Vassiliou

Bilgic E, Takao M, Kaneva P, Endo S, Takao T, Watanabe Y, McKendy KM, Feldman LS, Vassiliou MC. Development of a Model for the Acquisition and Assessment of Advanced Laparoscopic Suturing Skills using an Automated Device. 2017. [Under review]

- *Study conception and design:* Bilgic, Takao M, Kaneva, Watanabe, McKendy, Feldman, Vassiliou
- *Data acquisition:* Bilgic, Takao M, Endo, Takao T, Kaneva
- *Analysis and interpretation of data:* Bilgic, Kaneva
- *Drafting of manuscript:* Bilgic, Kaneva, Vassiliou
- *Critical revision:* Bilgic, Takao M, Kaneva, Endo, Takao T, Watanabe, McKendy, Feldman, Vassiliou

Bilgic E, Endo S, Lebedeva E, Takao M, Mckendy KM, Watanabe Y, Feldman LS, Vassiliou MC. A Comprehensive Review of Assessment Tools for Laparoscopic Suturing. [Under review]

- *Study conception and design:* Bilgic, Endo, Watanabe, Vassiliou
- *Data acquisition:* Bilgic, Lebedeva
- *Analysis and interpretation of data:* Bilgic, Endo, Takao, McKendy, Vassiliou
- *Drafting of manuscript:* Bilgic, Vassiliou
- *Critical revision:* Bilgic, Endo, Takao, McKendy, Watanabe, Feldman, Vassiliou

Bilgic E, Watanabe Y, McKendy KM, Munshi A, Ito YM, Fried GM, Feldman LS, Vassiliou MC. Reliable Assessment of Operative Performance. *American Journal of Surgery*. 2016;211(2):426-430.

- *Study conception and design:* Bilgic, Watanabe, Vassiliou
- *Data acquisition:* Bilgic, Watanabe, Munshi
- *Analysis and interpretation of data:* Bilgic, Watanabe, Ito
- *Drafting of manuscript:* Bilgic, Watanabe, Vassiliou
- *Critical revision:* Bilgic, Watanabe, McKendy, Munshi, Fried, Feldman, Vassiliou

STATEMENT OF ORIGINALITY

The work presented in this thesis represents original contributions to the understanding of assessment of laparoscopic suturing skills in simulation and operative settings.

In spite of the support and contribution of my supervisor, co-authors, and supervisory committee members, the work presented is original and novel.

STATEMENT OF SUPPORT

This thesis was supported by the Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, which is funded by an unrestricted educational grant from Covidien, and the Henry K.M. DeKuyper Education Centre, which is funded by the DeKuyper family and the Montreal General Hospital Foundation. They have provided the workspace and support necessary in terms of equipment and help from research assistants for the completion of my research. I have also received Entrance Fellowships and Tuition Assistance from the Graduate Excellence Fellowship Fund.

LIST OF TABLES AND FIGURES

assess procedures that required laparoscopic suturing in the operating room (e.g. laparoscopic nissen fundoplication).

- **Figure 1:** Study identification and selection flow chart.
- **Appendix 4.2.1:** MEDLINE search strategy.
- **Appendix 4.2.2:** Types of suturing and knot tying that were used with the assessment tools.
- **Appendix 4.2.3:** Summary of the validity evidence for assessment tools used in the simulation setting.
- **Appendix 4.2.4:** Details of the validity evidence of each assessment tool.
- **Appendix 4.2.5:** Guideline for the selection of an assessment tool that was used in the simulation setting.

Chapter 4.3 – Reliable Assessment of Operative Performance

- **Table 1:** Procedures included in the study.
- **Table 2:** Percent effect of each factor and their interactions on the performance score of the surgical trainees using GOALS.
- **Figure 1:** Reliability of number of assessments per trainee assessed by a single attending surgeon using D study.

ABSTRACT

The laparoscopic approach is currently a routine part of practice in general surgery with advanced laparoscopic procedures requiring technical skills, such as laparoscopic suturing (LS), being done more frequently than in the past. Due to increasing concerns for patient safety, the current training paradigm focuses on learning outside of the clinical setting through simulation as an adjunct to learning in the operating room (OR). This allows trainees to gain certain skills and improves the learning curve for operative performance. In order to track skill gain and make sure that trainees are competent in performing LS, performance assessment in the OR and simulation settings plays a vital role. This thesis investigates the accuracy of assessment platforms of fundamental and advanced LS skills in the simulation setting and the OR.

First, we conduct a scoping review to identify simulation platforms that have been developed to assess LS skills. Our results show that most platforms target basic LS skills, such as performance of one suture using intracorporeal knot-tying, or they use ex-vivo models, which are labor-intensive and not cost-effective. However, various needs assessments (conducted to determine gaps in current knowledge or resources available in a certain domain) identify that apart from needing strong training in basic LS skills, there is also a clear need for simulation platforms that target more advanced LS skills that are required in the clinical setting. These advanced LS skills include suturing under tension, suturing in tighter spaces, performing bowel anastamoses, and using automated-suturing devices. Therefore, we decided to investigate and develop simulation platforms that target the need for a more comprehensive assessment of LS, including basic and advanced skills.

Most educators already use an existing simulation program called Fundamentals of Laparoscopic Surgery (FLS), which has a didactic and simulation component for training and

assessment of basic knowledge and skills. The skills part of this program, when used with a proficiency curriculum, has been shown to improve fundamental laparoscopic skills, including LS. Some educational theories propose that new knowledge is developed from previous knowledge. Starting from fundamentals and moving to advanced LS skills allows trainees to grow their mental frameworks by incorporating previous and new experiences. To follow these theories while developing a comprehensive assessment platform for LS, we establish additional validity evidence for FLS and previously developed advanced LS tasks using free-needles. Additionally, we develop and provide validity evidence for an advanced LS task for device-assisted suturing.

The ultimate goal of simulation training is to ensure that skill gain transfers to the operative setting, and assessment allows us to measure the competence, or skill gain, of trainees. Therefore, it is important to understand how to best assess LS skills in the OR, which includes the selection and implementation of an assessment platform, such as an assessment tool. First, we complete a thorough literature review to identify assessment tools that have been used to assess skill in the OR, for both LS and procedures that require LS. This allows us to determine the validity evidence that is available for the tools and for which assessment conditions they were used. Among the tools with moderate evidence, the one with the highest evidence is for assessment of procedures that require LS by direct observation by the attending surgeon using a tool called Global Operative Assessment of Laparoscopic Skills (GOALS) that assesses generic laparoscopic skills (therefore it can be used to assess any laparoscopic procedure).

Finally, in order to understand how to implement a tool in the OR, we use GOALS. Operative performance can be impacted by various factors that are not related to the trainee performance, such as raters and cases (includes case difficulty and procedure type). Therefore,

we determine the impact of the external factors on the assessment score of the trainees. In addition, 1 assessment per trainee may not be sufficient for accurate representation of their skill level. Therefore, we also determine the number of assessments we will need per trainee to minimize the effect of external factors and to have an accurate representation of trainee performance. We find that, apart from the performance of the trainee, the way a trainee performs from case to case, independent of the rater, had a significant impact (32%) on the assessment score. In order to minimize the effect of the cases on the score, we find that assessment in at least 3 cases is required per trainee.

In conclusion, we have investigated assessment methods of laparoscopic suturing skills of trainees in simulation and clinical settings. This thesis presents strategies to improve assessment of trainees within training programs through the use of bench-top simulators and intraoperative assessment tools.

ABRÉGÉ

L'approche laparoscopique fait actuellement partie de la pratique courante en chirurgie générale avec des procédures laparoscopiques avancées nécessitant des compétences techniques, telles que la suture laparoscopique (SL), plus fréquentes que dans le passé. En raison des préoccupations croissantes pour la sécurité des patients, le paradigme de la formation actuelle se concentre sur l'apprentissage à l'extérieur du milieu clinique par la simulation en tant que complément à l'apprentissage en salle d'opération (SO). Cela permet aux stagiaires d'acquérir certaines compétences et d'améliorer la courbe d'apprentissage pour la performance opérationnelle. Afin de suivre le gain de compétences et de s'assurer que les stagiaires sont compétents dans la réalisation de SL, l'évaluation de la performance dans la SO et les paramètres de simulation jouent un rôle essentiel. Cette thèse étudie l'exactitude des plateformes d'évaluation des compétences SL fondamentales et avancées dans le cadre de la simulation et de la SO.

Premièrement, nous procédons à une revue de la portée afin d'identifier les plateformes de simulation qui ont été développées pour évaluer les compétences SL. Nos résultats montrent que la plupart des plateformes ciblent les compétences SL de base, telles que la performance d'une suture utilisant des nœuds intracorporels, ou qu'elles utilisent des modèles ex vivo, qui demandent beaucoup de main-d'œuvre et ne sont pas rentables. Cependant, diverses évaluations des besoins (menées pour déterminer les lacunes dans les connaissances actuelles ou les ressources disponibles dans un certain domaine) indiquent qu'en plus d'avoir besoin d'une solide formation en compétences de base en SL, il existe un besoin évident de plateformes de simulation dans le cadre clinique. Ces compétences avancées en SL comprennent la suture sous tension, la suture dans des espaces plus étroits, l'anesthésie intestinale et l'utilisation de dispositifs de suture automatisés. Par conséquent, nous avons décidé d'étudier et de développer

des plateformes de simulation qui ciblent la nécessité d'une évaluation plus complète de SL, y compris les compétences de base et avancées.

La plupart des éducateurs utilisent déjà un programme de simulation intitulé « Fundamentals of Laparoscopic Surgery » (FLS), qui comporte une composante didactique et de simulation pour la formation et l'évaluation des compétences de base. Ce programme a été établi pour améliorer les compétences laparoscopiques fondamentales, y compris SL. Certaines théories éducatives proposent que toute la connaissance est développée à partir des connaissances antérieures. Partir des fondamentaux et passer à des compétences avancées en SL permet aux stagiaires de développer leur cadre mental en intégrant des expériences antérieures et nouvelles. Afin de suivre ces théories tout en développant une plateforme d'évaluation complète pour SL, nous établissons des preuves de validité supplémentaires pour FLS et des tâches SL avancées précédemment développées pour l'aiguille libre. De plus, nous développons et fournissons des preuves de validité pour une tâche SL avancée pour la suture assistée par dispositif.

Le but ultime de la formation par simulation est de faire en sorte que les gains de compétences soient transférés au contexte opérationnel, et l'évaluation nous permet de mesurer la compétence ou le gain de compétences des stagiaires. Par conséquent, il est important de comprendre comment évaluer plus mieux les compétences en SL dans la SO, ce qui comprend la sélection et la mise en œuvre d'une plateforme d'évaluation, telle qu'un outil d'évaluation. Premièrement, nous effectuons une analyse documentaire approfondie afin d'identifier les outils d'évaluation qui ont été utilisés pour évaluer les compétences dans la salle d'opération, à la fois pour le SL et les procédures nécessitant une SL. Cela nous permet de déterminer les preuves de validité disponibles pour les outils et pour quelles conditions d'évaluation elles ont été utilisées.

Parmi les outils avec des preuves modérées, celui avec les preuves les plus élevées est pour l'évaluation des procédures qui nécessitent SL par observation directe par le chirurgien traitant en utilisant un outil appelé l'évaluation globale opératoire des compétences laparoscopiques (GOALS) qui évalue les compétences laparoscopiques génériques (donc il peut être utilisé pour évaluer toute procédure laparoscopique).

Enfin, pour comprendre comment implémenter un outil dans la SO, nous utilisons des GOALS. La performance opérationnelle peut être affectée par divers facteurs qui ne sont pas liés à la performance du stagiaire, tels que les évaluateurs et les cas (comprend la difficulté du cas et le type de procédure). Par conséquent, nous déterminons l'impact des facteurs externes sur le score d'évaluation des stagiaires. En outre, une évaluation par un stagiaire peut ne pas être suffisante pour une représentation précise de leur niveau de compétence. De ce fait, nous déterminons également le nombre d'évaluations dont nous aurons besoin par un stagiaire pour minimiser l'effet des facteurs externes et avoir une représentation précise de la performance du stagiaire. Nous constatons que, mis à part la performance d'un stagiaire, la façon dont le stagiaire se comporte d'un cas à l'autre, indépendamment de l'évaluateur, a eu un impact significatif (32%) sur le score d'évaluation. Afin de minimiser l'effet des cas sur le score, nous trouvons qu'une évaluation dans au moins 3 cas est requise par un stagiaire.

En conclusion, nous avons étudié les méthodes d'évaluation des compétences de suture laparoscopique des stagiaires dans la simulation et les paramètres cliniques. Cette thèse présente des stratégies pour améliorer l'évaluation des stagiaires dans les programmes de formation à l'aide de simulateurs de paillasse et d'outils d'évaluation peropératoire.

CHAPTER 1: INTRODUCTION

1.1 Background

Over the years, surgical education has evolved with more research focusing on how to best train surgeons. In the past, most surgical training happened during clinical encounters and by apprenticeship. However, in more recent years, this form of training brought challenges to surgical educators due to concerns regarding a lack of consistency in teaching methods, a decrease in resident work hours, and most importantly, concerns regarding patient safety and ethics[1]. Trainees need a platform to acquire skills outside of the clinical setting, so that when they operate on patients, they are more safe and efficient. This is where simulation comes into play. For instance, Seymour et al. conducted a study where residents were either trained in a virtual reality (VR) simulator or they were controls (conventional training)[2]. When the residents were assessed post-training performing gallbladder dissection in the operating room (OR), they found that simulation-trained residents were faster and 6 times less likely to conduct errors, such as liver injury, burning non-target tissue, etc. Antosh et al. found similar results in terms of improvement of skill in their study. They found that training with the Fundamentals of Laparoscopic Surgery (FLS) tasks improves suturing skills in the OR for gynecology residents[3]. These studies are just some examples that showcase how simulation training improves the technical skills of residents. In addition, the Michigan Bariatric Surgery Collaborative team conducted a study that asked surgeons to submit a video of them performing laparoscopic gastric bypass. They assessed the surgeons' video-recorded performances and evaluated whether there is a relationship between post-operative complication and mortality rates and surgeon technical skills[4]. The technical skills were assessed through an intraoperative

assessment tool called Objective Structured Assessment of Technical Skills (OSATS), which has 5 items and each item is rated on a scale of 1-5, 1 being lower skill level. They found that surgeons with lower scores were associated with having higher complication and mortality rates compared to surgeons with higher scores. This suggests that there is a link between patient safety and the technical skills of surgeons.

In order to improve the training, and hence address concerns of patient safety and ethics, programs have shifted towards 'Competency-based Medical Education' (CBME), which outlines core competencies in which trainees must obtain proficiency[5, 6]. This form of training focuses on trainee outcomes with regards to ability, skill gain, and proficiency in a certain domain, and the curriculum is learner-centered and can be individualized based on the needs of each student [7]. Within this paradigm, a significant amount of training is done in the simulation setting, whereby trainees can practice a skill repeatedly without hindering patient safety[1].

Assessment is vital and ensures that trainees are competent in various domains, tracks trainee progress, and allows programs to make decisions about trainee proficiency[8]. Traditionally, performance assessment was based on case numbers and subjective evaluations at the end of rotations among other methods. However, with CBME, the shift has been made towards more objective assessments of observable behaviors in both clinical and simulation settings, in order to accurately measure trainee performance using various instruments with established validity evidence[9-11].

One area that uses CBME is the training of laparoscopic skills. In the early 1990s, laparoscopy started to gain popularity and now, it is commonly used in a variety of different surgical procedures[12]. Due to the differences in knowledge and skills required to perform laparoscopy versus open cases, a need has emerged with regards to how to best assess

laparoscopic skills, with one focus being on laparoscopic suturing (LS)[13]. In order to address this need, various assessment platforms have been developed for LS in both simulation and clinical settings, whereby trainee competence can be accurately assessed. Assessment platform refers to the task being used to measure skill and the assessment metrics that are used to score the performances (including assessment tools, motion analysis etc).

Laparoscopic skill assessment in simulation

Simulation plays an important role in the current training for laparoscopic surgery, including laparoscopic suturing(LS) [14-16]. One of the first platforms developed to teach and assess laparoscopic skills is the Fundamentals of Laparoscopic Surgery (FLS) program, which has online didactic modules and a skills curriculum, in addition to an assessment of both cognitive and manual skills[17]. The manual skills component measures basic laparoscopic technical skills, and includes tasks, such as Peg transfer, Pattern cutting, Ligating loop, Extra and intra-corporeal suturing and knot tying. The FLS program is currently used widely, and there is plenty of evidence to suggest that FLS, among other simulation training platforms, enhances skill acquisition, and correlates with better performance in the OR[2, 18-20]. For example, Sroka et al. conducted a randomized control trial to determine if training with the FLS proficiency-based curriculum improved skill in the OR, and they found that trainee performance in the OR improved in comparison to a control group who did not go through the curriculum[18].

Laparoscopic surgery is currently a routine part of practice in general surgery, with advanced laparoscopic procedures that require LS being performed commonly[12]. Due to this change in practice routines, trainees are being exposed to advanced laparoscopic cases earlier and more often in their training, which provides trainees with increased opportunities to learn

advanced laparoscopic skills if they are appropriately prepared.

Needs assessments are often conducted to determine if there are gaps between what is currently available and what needs to be available in a certain domain, so that any gaps can be addressed through resource development. Various needs assessments have been conducted to determine laparoscopic skills missing from current simulation platforms that should be taught and assessed. Mattar et al. conducted a survey of program directors (PD) and identified that for technical skills, 56% of the PDs did not feel that graduated residents were proficient at performing laparoscopic suturing (LS)[21]. Additionally, from conducting interviews of general surgery residents and surgeons, Enani et al. identified a major obstacle to mastering LS skills: there are limited simulation platforms that allow trainees to practice LS skills that are more complex and currently necessary. These complex LS skills include both free-needle and device-assisted suturing[22]. Finally, a survey of fellows and PDs by Nepomnayshy et al. identified a need for an advanced laparoscopic surgery simulation curriculum within residency programs, with one of the main focuses being LS[23]. They have also determined that the training of LS skills should focus on anastomosis, suturing under tension, suturing in tighter spaces, backhand suturing, suturing with more realistic camera angles, bimanual dexterity, improving use of non-dominant hand, and tissue handling (using fragile tissue). These three needs assessments identified that, with the advancement of laparoscopic procedures in practice, LS is a skill where there is a gap between what is currently taught and assessed (basic LS skills) versus skills that are required in the clinical setting. Platforms, such as FLS, that teach and assess basic LS skills do not focus on the more advanced aspects of suturing identified as important by Nepomnayshy et al. In addition to these findings, a recent review by Lim et al. regarding the economic and clinical effects of LS and its impact in the adaptation of laparoscopy concluded that LS still

remains as a complex skill, and its complexity stops some surgeons from adapting advanced laparoscopic procedures into their practice[24].

One theory of motor learning has been proposed by Fitts and Posner. They suggest that acquiring a skill like LS happens in 3 phases: cognitive (the resident has the knowledge of how to suture; how to use the instruments, tie a knot, etc.), associative (with repeated practice, LS knowledge becomes action and residents can perform LS more fluidly), and autonomous (with repeated practice, LS performance becomes smoother where trainee performs without thinking about how to suture; the skill became automatic)[25]. One main difference between an expert and a novice performing a task is that when an expert performs a task, they are often in a state of automaticity. This means that the skill has become autonomous for the expert and they do not have to use much of their cognitive capacity (the total amount of information that they can retain) towards completing that task. When the novice performs the same task, this is the opposite[26].

Automaticity is reached through repeated practice, and simulation training plays an essential role as it allows novices to learn LS through repeated practice without concerns over patient safety, time constraints, or cost that would limit training in the clinical setting. When the trainee reaches that state, they have created space in their cognitive capacity to focus on other aspects of LS in the clinical setting that might be harder to simulate, such as error prevention, or interferences in the OR environment (individuals involved in the case, problems with equipment, etc.). This way, trainees can learn LS in the simulation setting, and use the valuable time they have in the OR to apply what they learnt in simulation and build on their experience. Consequently, simulation training provides a valuable platform for trainees to practice

repeatedly, make mistakes, and remediate their mistakes, allowing them to come to the OR with prior knowledge and skill that they can build upon.

Therefore, it is critical to develop platforms to ameliorate the adequacy of current training with regards to providing resources to assess advanced LS skills while maintaining a strong fundamental skills assessment, which will be addressed in the first 2 chapters of this thesis[14]. Among many of the adult learning theories, constructivism supports the idea that new knowledge is developed from previous knowledge, and each trainee learns by incorporating their previous knowledge and interactions with new experiences and individuals[25]. From this point of view, starting from the fundamentals and then moving to advanced LS skills allows trainees to develop their mental framework for LS by incorporating previous and new experiences. If a trainee is not competent with the fundamental steps required to perform LS (from how to hold/orient the needle to how to tie a knot), then they are not going to have previous knowledge to build upon and be able to perform suturing in more complex environments where the angle is different or the tissue is under tension[27].

Laparoscopic skill assessment in the operating room

Simulation training is certainly very important for trainees to gain a skill, such as LS. However, the next step after simulation is to determine if trainees are improving in the clinical setting and if they are ready for the next stage of their training. This is crucial as the ultimate goal for trainees is to use simulation platforms to improve their skills to become competent in the operating room, which, in the end, will improve patient safety[28]. For assessment in the operative setting, there are various ways to assess trainees. These ways include computer-assisted technologies, such as eye-tracking, motion analysis, and usage of assessment tools to

assess various aspects of competence[29, 30]. There is emerging evidence to suggest that assessment tools provide a low-cost assessment platform where objective assessments of trainee performance can be completed for a variety of skill-sets[31, 32]. One of the examples where programs have gone beyond the research setting and actually implemented an assessment tool into the clinical setting has come from the American Board of Surgery (ABS). The ABS has made it mandatory for institutions in the United States of America to assess trainee skill in the clinical setting for general surgery residents using the Operative Performance Rating System (OPRS)[33]. This was initiated because operative assessments are important to ensure that all trainees have a certain level of knowledge and skill when graduating.

Assessment requires a need to clearly define the reason for assessment and what exactly is being assessed (feedback or decision making, technical skills or knowledge, etc.) and the measures that will be used for that assessment (e.g. usage of assessment tools). The point here is that through assessments, we are gathering information about someones' performance and inferring about their skill level. In order to accurately draw conclusions from the measurements, researchers need to provide validity evidence. Validity evidence answers the question of 'does the tool measure what we are intending to measure', which is evidence for the interpretation of the assessment score, rather than evidence for the tool itself[11]. Therefore, the evidence that we collect allows us to assign meaningful interpretations to the assessment scores in order to accurately reflect the actual skill level of the trainees. Validity evidence is context specific; an assessment could be done for decision making or to provide feedback for improvement. Assessment could also be done through direct observation or video-assessment, and the rater could be the attending surgeon or an observer. Validity has to be provided in each context separately; an assessment that has evidence for high-stakes video-assessment cannot be

generalized to any other context. One of the ways in which researchers can build an argument regarding the accurate interpretation of scores, which, as stated above, depends on the context of assessment, is through the validity framework by Messick. This framework includes 5 sources of validity: content, response process, internal structure, relations to other variables, and consequences[34]. Each of these sources is an important component of validity and should be addressed to have a complete understanding of the validity evidence surrounding an assessment score.

Even though operative assessment is an area of great interest currently, there are still limitations when it comes to assessment of LS skills and implementation in training programs[35]. First, it becomes challenging for surgical educators to reach consensus when selecting an assessment tool for their purposes and assessment conditions. For instance, assessment tools could be specific to LS, an advanced procedure that requires LS, or generic to laparoscopy. The trainees could be assessed through direct observation or video recording, and they could be assessed by an observer, the attending surgeon of the case, or trainees themselves could be the raters. Validity evidence has to be provided for each assessment condition separately; evidence for one condition cannot be generalized to another[34]. Therefore, when programs are in the process of selecting an assessment tool for implementation, they have to consider the validity evidence available, and in what context that evidence was provided.

Another limitation is that apart from selecting an assessment tool for LS, it is important to reach a consensus regarding how to actually implement an assessment tool. The OR is a complex environment where there are a lot of factors that could affect the performance of the trainee and hence their assessment score, which are independent of the trainee skill level[36]. These factors could include rater bias, difficulty of the case, or differences in the procedures. Therefore,

programs have to find ways to minimize the effect of external factors on the assessment score to properly assess trainees and ensure that their score truly reflects their performance.

In the end, it is crucial for surgical educators to identify an assessment tool for LS that fits the context of their assessment, has validity evidence to support its use in that context, and determine how to implement that assessment tool into their program, so that performance assessments can be done appropriately in the clinical setting. Addressing these limitations will be the other focus of this thesis.

1.2 Thesis Objectives

The specific objectives of this thesis are:

1. To identify simulation platforms that have been developed for assessment of laparoscopic suturing (LS) skills;
2. To evaluate the current trends of the Fundamentals of Laparoscopic Surgery exam that includes the assessment of basic laparoscopic suturing skills;
3. To develop and provide validity evidence for advanced laparoscopic suturing tasks as measures of advanced LS skills, including both free-needle and device-assisted suturing;
4. To develop a guideline for surgical educators to aid in the selection of an assessment tool used to assess LS specifically or to assess procedures that require LS in the operating room;
5. To determine the impact of raters, cases, and trainee performance on the assessment scores of the trainees in the operating room using the Global Operative Assessment of Laparoscopic Skills (GOALS) assessment tool, and to determine the number of

assessments required to have reliable assessment scores that truly reflect trainee performance.

To achieve our aims, first, we did a literature review of simulation platforms available for assessment of LS. Then, we evaluated an existing simulation program called FLS, which has been shown to accurately assess fundamental knowledge and skills required in laparoscopy (including LS). Subsequently, based on previous needs assessments, we addressed the need for simulation platforms that better reflect the complexities of the LS skills required in the OR (beyond the fundamentals that are targeted in FLS and other training platforms) by developing and providing validity evidence for advanced LS tasks as measures of LS skills. Even though simulation training is important, that gain of skill obtained must be reflected in the clinical setting, and hence, assessment of performance in the operating room (OR) becomes crucial. Therefore, next, we shifted our focus to clinical assessment. We conducted an in-depth review of the literature in order to identify assessment tools that have been used to assess LS skills as well as procedures that require LS in the OR, and determine the validity evidence surrounding the instruments for various assessment conditions. After performing the review, we selected one of the assessment tools, and determined the effects of various factors on the assessment score, and how we can minimize these effects in order to have consistent scores that reflect the true operative-performance of the trainees.

CHAPTER 2: IDENTIFYING SIMULATION PLATFORMS FOR ASSESSMENT OF LAPAROSCOPIC SUTURING SKILLS

2.1 Pre-amble

Advanced laparoscopic procedures requiring suturing are now commonly performed. Laparoscopic suturing (LS), however, still remains as a skill that most trainees are not proficient at. In this chapter, the literature is reviewed in order to identify simulation platforms that are available for assessment of LS skills, including the type of simulation and materials used, along with the type of LS that is targeted. This is the first step in allowing us to address the need for LS simulation tasks that better reflect complexities in the operating room.

2.2 Simulation Platforms to Assess Laparoscopic Suturing: a Scoping Review

Running Head: Simulation Platforms of Laparoscopic Suturing

Authors:
Elif Bilgic[1] BSc, Motaz Alyafi[1] MD, Tara Landry MLIS[2], Melina C Vassiliou[1] MD

Institutions and Affiliations:
1. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada
2. Montreal General Hospital Medical Library, McGill University Health Centre, Montréal, QC, Canada

Conflicts of Interest and Sources of Funding:
All authors have no relevant conflicts of interests to disclose. The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Medtronic Canada. There was no funding or financial support for this work.

Corresponding author
Melina C. Vassiliou, MD
Associate Professor of Surgery
McGill University Health Centre
Montreal General Hospital
1650 Cedar Avenue, L9-313
Montreal, QC, H3G 1A4
melina.vassiliou@mcgill.ca

Abstract:

Background:

Laparoscopic suturing (LS) has become a common technique used in a variety of advanced laparoscopic procedures. However, various needs assessments identified that LS is a challenging skill to master, and we need to make sure that trainees are competent in performing LS at the end of their training. The purpose of this review is to identify simulation platforms available for assessment of LS skills, and determine the characteristics of the platforms and the LS skills that are targeted.

Methods:

A scoping review was conducted between January 1997 and October 2017 for full-text articles. The search was done in various databases. Only articles written in English or French were included. Additional studies were identified through reference lists. The search terms included "laparoscopic suturing" and "competence."

Results:

Fifty-one studies were selected. The majority of the simulation platforms were box trainers with inanimate tissue, and targeted basic, 1 suture intracorporeal knot-tying techniques. Most of the validation came from internal structure (rater reliability) and relationship to other variables (compare training levels/case experience, compare various metrics). Consequences was not addressed in any of the studies.

Conclusion:

We identified many types of simulation platforms that were used for assessing LS skills, with most being for assessment of basic skills. Platforms assessing the competence of trainees for advanced LS skills was limited. Therefore, future research should focus on development of LS tasks that better reflect the needs of the trainees.

Key words: Laparoscopy, Suturing, Simulation, Assessment

Background

Laparoscopic suturing (LS) is a skill needed in a variety of advanced laparoscopic procedures that have become a routine part of practice, where LS is used for bowel anastamosis, closure of hiatel defects, handling complications, and other procedures. This skill has been identified as one of the more challenging skills for surgical trainees to master[37, 38]. Needs assessments by Nepomnashy et al and Enani et al identified a gap between LS skills needed in the operating room (OR) and LS skills targeted by various simulation platforms[22, 23]. The main gaps were identified for suturing under tension, suturing in tighter spaces, performing bowel anastamosis, backhand suturing, and suturing using automated devices. These findings were in conjunction with a survey by Mattar et al where they found that more than half of the program directors (PD) did not think graduated residents had enough skills to perform LS in the OR[21]. Together, these findings all suggest that there is a need to improve the training of the residents when it comes to LS. In order to track trainee progress and make sure that they are competent in performing LS, we need to be able to assess their skill. Assessment allows us to ensure that when residents graduate, they are competent in LS. The purpose of our scoping review is to identify simulation platforms available for assessment of LS skills, and determine the characteristics of the platforms and the LS skills that are targeted.

Materials and Methods

Search strategy

We performed a scoping review between January 1997 and October 2017 for full-text articles. Search strategies were developed with a librarian (T.L.). The search was done in MEDLINE, Embase, CENTRAL, CDSR, and PubMed. MEDLINE search strategy (Appendix 2.2.1) was applied to all databases. Only articles written in English or French were included.

Additional studies were identified through reference lists. The search terms included "laparoscopic suturing" and "competence."

Study selection

Included studies reported data on development and/or validation of simulation tasks for assessment of LS. Studies were excluded if they (1) only included medical students, (2) were assessment of whole procedures, (3) were part of a program for urology or gynecology or involved LS specific for those specialties, or (4) were educational intervention.

Data extraction

Two independent reviewers (E.B., M.A.) conducted the screening. Extracted information included type of simulator (box trainer, augmented reality, virtual reality), task (in-vivo, ex-vivo, inanimate), suturing (intracorporeal (ICK), extracorporeal (ECK), continuous, interrupted, hand-sewn, device assisted), scoring metrics, and sources of validity.

Validity

Validity refers to the evidence surrounding a simulation task that measures LS skills. There are 5 sources of validity evaluated: content (can the simulation tasks measure suturing skills), response process (can the scoring be done accurately), internal structure (are the scores consistent), relations to other variables (do the task scores correlate with other assessments or differentiate between training levels), and consequences (what are the implications of incorporating the task for assessment into the training programs)[39, 40].

Results

Simulation platforms

Through our search, we included 51 studies for data analysis (Figure 1)[26, 41-90]. Some studies used multiple suturing techniques and metrics, which is why the numbers do not add up

to 51. Among them, 38 used the box trainer (32 inanimate, 5 ex-vivo, 1 not specified(NS)), 10 augmented reality, 3 virtual reality, 0 *in vivo* methods, and 1 cadavers. The majority of the suturing was done using basic ICK techniques (interrupted, 1 suture ICK), similar to the suturing done in the Fundamentals of Laparoscopic Surgery (FLS) ICK task. For knot-tying, 45 studies used ICK, 3 ECK, and 7 no knot-tying. For suturing, 3 studies used continuous suturing, 44 interrupted, 47 hand-sewn, and 1 device assisted. Few studies assessed advanced LS skills. Of those studies, 2 assessed anastamosis techniques with continuous suturing using a porcine intestine model, 1 assessed the anastamosis technique with interrupted suturing using a synthetic intestine model, and 1 assessed ICK in a difficult location on a lamb liver (deep suturing). There was also one major vessel injury (MVI) model where trainees had to perform suturing of any kind to stop the bleeding of a synthetic tissue. In terms of metrics, 24 studies used multiple metrics, 0 used time alone, 6 time and error alone, 16 motion analysis alone, and 4 assessment tools alone. More details can be found in Appendix 2.2.2.

Validity evidence

The majority of the evidence came from the internal structure where they evaluated rater consistency, and the relationship to other variables where they evaluated differences between different training levels (Appendix 2.2.3). Four studies investigated evidence for content through expert opinion and no studies investigated response process or consequences.

Discussion

This review identified studies that developed and/or provided validity evidence for simulation tasks in the context of assessment of LS skills. The first point that we identified is that most suturing tasks were based on inanimate models, targeting basic LS skills such as 1 interrupted suture with intracorporeal knot-tying. One of the platforms most often used was the

FLS ICK task, or studies modifying the ICK task to fit their model. Regarding advanced LS skills, such as suturing under tension and bowel anastamosis, that Nepomnashy et al and Enani et al identified as needing simulation platforms, the only models available were ex vivo, with 1 synthetic intestine model[23]. Other identified skills, such as backhand suturing, suturing in tight spaces, and suturing using automated devices, had no simulation platforms associated with them (only 1 study included suturing with an automated device).

Metrics are an essential component to simulation since they allow us to objectively assess learner performance and make sure that learners are competent in a given domain, such as LS. Various metrics could have a role in high stakes assessment (measuring competence, decision-making) or low-stakes assessment (providing feedback), and the metrics have to be linked to the purpose of assessment[91, 92]. Among the studies identified, the majority of the metrics used for assessment were time and error, motional analysis, and assessment tool scores. Time and error metrics are easy to implement and there is plenty of data supporting their use for assessment. But if the assessment's purpose was to provide feedback, time and error metrics only target speed and the end product, which limits learning as the process taken to reach the end product is just as important as the end product itself and evaluating the process discourages formation of bad habits inappropriate to the clinical setting[93, 94]. Motion analysis is a combination of computer generated metrics. Although motion analysis removes the human factor and improves score consistency, interpreting the meaning behind the scores is not always clear; just because someone had a similar path length to experts does not mean that their end product is clinically sufficient (e.g. knots do not come off, there is no leak, etc.)[95, 96]. Finally, assessment tools can provide meaningful feedback and capture the process of how someone achieved the end product, but they require raters, which could be resource intensive[97]. None of the identified studies used

the assessments for high-stakes evaluation of skill level. They all provided evidence in the context of assessing competence without specifying the purpose of assessment. As stated above, having a clear purpose is important when choosing the measures of assessment. Overall, the majority of the time, metrics were a combination of various types. The reason for using different types of metrics could be that due to the complex nature of the LS skill (even when it is basic LS), educators are trying to capture a more complete picture of the trainee performance.

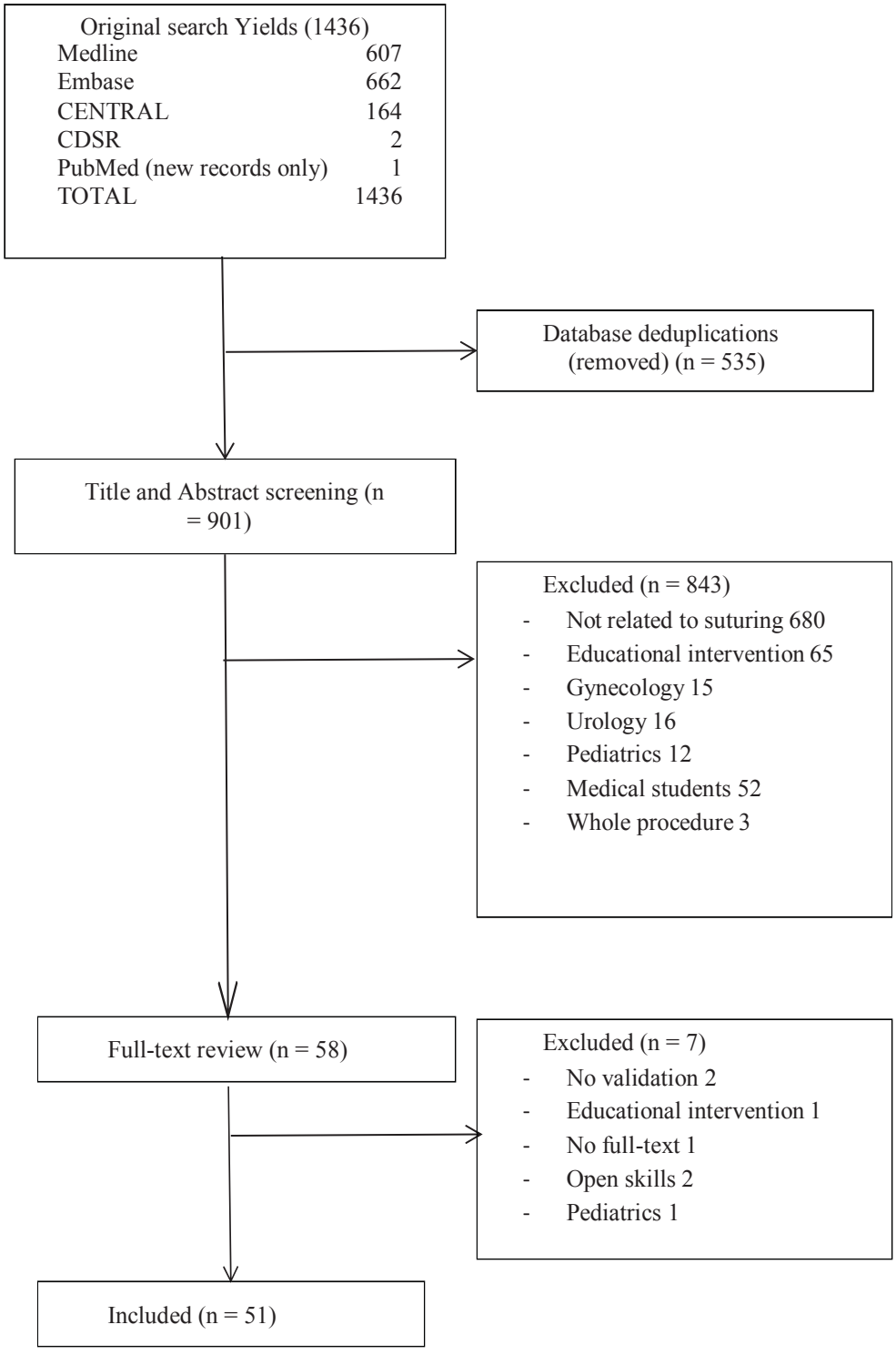
Regarding the validity evidence gathered for LS skills assessment, most studies investigated the internal structure through rater reliability and relationship to other variables by comparing scores of different training groups or correlation of various metrics, with minimal emphasis of the other 3 sources of validity. Within this validity framework, all sources add something important and the more sources of validity investigated, the more robust the evidence for the tools' potential to measure LS skills. To provide evidence for the content, the tasks could be developed with experts in the field in order to make sure that the skills assessed represent skills needed in the clinical setting. Furthermore, to provide evidence for the response process, steps could be taken to ensure that the scoring is done accurately, through rater training and clarification of what the various performance scores mean. Evidence for the consequences is not addressed much in the literature because it requires a longitudinal investigation of the implications. For example, if the assessor sets pass/fail standards for the assessment, the consequences involve what happens to the trainees who receive a fail grade and what steps are taken to make sure that they can pass. Above all, what we have to understand is that validity is not about the task, it is about the assessment and interpretation of the score trainees receive. This relates to the metrics of the tasks, since metrics construct the score[39]. Through the 5 sources, educators build an argument about the interpretation of the assessment depending on the purpose

of using the simulation task: high-stakes versus low-stakes assessment (validity of the interpretations that we made). This is important since evidence gathered in each of the 5 sources may differ depending on the purpose. Additionally, in this review, we did not investigate simulations that were used as educational intervention, which might also require different validity evidence. Therefore, establishing a context-dependent validity evidence is important so that educators can accurately reach a conclusion from the evidence gathered. As stated in the previous paragraph, even though all of the studies provided evidence in the context of assessing LS competence, they were not clear on the purpose of assessment.

Among the 51 studies that were analyzed, we identified a lot of different simulation tasks that were used for assessment purposes: box trainer, VR, and augmented reality. In addition, various types of metrics and suturing and knot-tying types were used. The variety of tasks and metrics illustrate a lack of consensus regarding the best way to incorporate simulation platforms to assess LS skills. However, it is also evident from our review that all platforms had varying degrees of validity evidence in the context of assessing LS skills. Therefore, it is more important to consider the limitations of each platform and choose a platform accordingly. For example, the cost associated with the platforms and the feasibility of obtaining the necessary apparatus may be limiting for some programs. VR is expensive, yet it allows easy assessment of trainees while performing a whole procedure without needing live animal models. Box trainers with ex-vivo tissue might increase the fidelity of the task, yet it is hard to preserve the tissues and they are more labor intensive. Box trainers with inanimate models such as penrose and fabrics are a much cheaper platform, which could explain why 73% of the analyzed studies used them, and there is ample evidence to suggest that they are an effective assessment platform.

In conclusion, we identified simulation platforms used for assessing LS skills. There were plenty of different platforms, yet platforms that could assess the competence of trainees for advanced LS skills identified by Nepomnayshy et al (suturing under tension, anastamosis, backhand suturing, suturing using an automated device, suturing in tight spaces) lacked representation. Only 4 studies assessed anastamosis and suturing in difficult locations, mostly using ex-vivo models which are more resource intensive. Therefore, there is a need for simulation platforms that can assess advanced LS skills, using low-cost and readily available materials so they are accessible to a wide range of training programs.

Figure 1: Study selection flow chart



CHAPTER 3: DEVELOPING AND PROVIDING VALIDITY EVIDENCE FOR FUNDAMENTAL AND ADVANCED LAPAROSCOPIC SUTURING TASKS IN SIMULATION FOR ASSESSMENT OF LAPAROSCOPIC SUTURING SKILLS

3.1 Pre-amble

The previous chapters have provided an understanding of the current simulation platforms available for assessing laparoscopic suturing (LS). Essential technical skills like LS have both fundamental and advanced components, and it is important for trainees to have a grasp of the fundamentals before moving on to the advanced skills. Therefore, incorporating both components allow for a more comprehensive assessment of skill.

One of the platforms identified in chapter 2 was Fundamentals of Laparoscopic Surgery (FLS), which as stated previously, includes a teaching and an assessment component for cognitive and manual skills, including LS. This program has plenty of evidence suggesting its effectiveness in assessing basic laparoscopic skills. Therefore, FLS still provides an essential component to the assessment of fundamental LS skills, which is why the first part of this chapter will explore the FLS exam in order to evaluate current trends and scoring practices.

At the same time, various needs assessments identified that when it comes to advanced LS skills, there is a lack of simulation platforms that capture the complexities of performing LS in the clinical setting, which will be addressed in the second part of this chapter. We are developing and/or providing validity evidence for advanced laparoscopic suturing tasks as measures of LS skills, for both free-needle and device-assisted suturing.

3.2 Trends in the Fundamentals of Laparoscopic Surgery ® (FLS) Certification Exam Over the Past 9 Years

Running Head: Fundamentals of Laparoscopic Surgery ® (FLS) Trends of 9 Years

Authors:
Elif Bilgic[1] BSc, Pepa Kaneva[1] MSc, Allan Okrainec[2] MD, E Matthew Ritter[3] MD, Steven D Schwaitzberg[4] MD, Melina C Vassiliou[1] MD

Institutions and Affiliations:
1. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada
2. Department of Surgery, University Health Network, University of Toronto, Toronto, ON, Canada
3. Department of Surgery at Uniformed Services University of the Health Sciences and Walter Reed National Military Medical Center, Bethesda, MD, USA
4. Department of Surgery, University at Buffalo, Jacobs School of Medicine and Biomedical Sciences, Buffalo, NY, USA

Conflicts of Interest and Sources of Funding:
This is a quality initiative of the SAGES FLS committee. Elif Bilgic, Pepa Kaneva, and Drs Allan Okrainec, E. Matthew Ritter, Steven D. Schwaitzberg and Melina C. Vassiliou have no relevant conflicts of interest or financial ties to disclose. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Covidien Canada. There was no funding or financial support for this work.

Corresponding author
Melina C. Vassiliou, MD
Associate Professor of Surgery
McGill University Health Centre
Montreal General Hospital
1650 Cedar Avenue, L9-313
Montreal, QC, H3G 1A4
melina.vassiliou@mcgill.ca

Abstract:

Background:

The Fundamentals of Laparoscopic Surgery ® (FLS) certification exam assesses both cognitive and manual skills, and has been administered for over a decade. The purpose of this study is to report results over the past 9 years of testing in order to identify trends over time and evaluate the need to update scoring practices. This is a quality initiative of the SAGES FLS committee.

Methods:

A representative sample of FLS exam data from 2008- 2016 was analyzed. The de-identified data included demographics and scores for the cognitive and manual tests. Standard descriptive statistics were used to compare trends over the years, training levels, and to assess the pass/fail rate.

Results:

A total of 7232 FLS tests were analyzed (64% male, 6.4% junior(PGY 1-2), 84% senior(PGY3-5), 2.8% fellows(PGY6), and 6.7% attending surgeons(PGY7)). Specialties included 93% general surgery(GS), 6.2% gynecology, and 0.9% urology. The Pearson correlation between cognitive and manual scores was 0.09. For the cognitive exam, there was an increase in scores over the years, and the most junior residents scored lowest. For the manual skills, there were marginal differences in scores over the years, and junior residents scored highest. The odds ratio of PGY3+ passing was 1.8 (CI 1.2-2.8) times higher than a PGY1-2. The internal consistency between tasks on the manual skills exam was 0.73. If any one of the tasks was removed, the Cronbach's alpha dropped to between 0.65 -0.71, depending on the task being removed.

Conclusion:

The cognitive and manual components of FLS test different aspects of laparoscopy and demonstrate evidence for reliability and validity. More experienced trainees have a higher likelihood of passing the exam and tend to perform better on the cognitive skills. Each component of the manual skills contributes to the exam and should continue to be part of the test.

Keywords: Fundamentals of Laparoscopic Surgery, Laparoscopy, Simulation, Skill Assessment

Background

Laparoscopic surgery is used worldwide to perform a variety of surgical procedures. The knowledge and skills specific to laparoscopy differ from those needed for open procedures. The Society of Gastrointestinal and Endoscopic Surgeons (SAGES) created the Fundamentals of Laparoscopic Surgery ® (FLS) program in the late 1990's in order to provide a consistent curriculum and validated assessment of cognitive and manual skills[17].

FLS testing has been administered for over a decade and includes didactic educational material and simulation-based training for technical skills, along with rigorous assessment of both the cognitive and manual skills[98]. The didactic portion includes preoperative, intraoperative and postoperative considerations specific to laparoscopy, and not specific to any surgical subspecialty. The technical skills portion was based on 5 of the original 7 McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) tasks[91]. There is ample evidence for the validity of the FLS examination as a measure of the basic knowledge and skills required for laparoscopic surgery[93, 99-101].

The first FLS testing was done in 2004 and the results from the first 5 years have been reported[102]. The purpose of this study is to report FLS test results from the past 9 years in order to describe current trends and evaluate the need to update scoring practices. This is a quality initiative of the SAGES FLS committee with the goal of keeping the exam current and relevant to the changing clinical environment.

Methods

The FLS cognitive exam consists of multiple-choice questions and the manual skills exam consists of 5 tasks that are completed in a box trainer. The details have been described elsewhere[13, 17, 103]. A representative sample of FLS exam data that was prospectively collected from 2008- 2016 was analyzed. The de-identified data included demographics such as age, postgraduate year (PGY), specialty, and scores for the cognitive and manual tests.

Pearson correlation was used to assess associations between cognitive and manual skill scores. Internal consistency of the manual skill scores was calculated using Cronbach's alpha. Generalized linear models (GENMOD) (SAS 9.4) were used to assess pass/fail rates and to compare trends over time and experience levels. Identity or logit link functions were used to fit linear or logistic regression models. Statistical significance was <0.05. Tukey-Kramer adjustments were used for multiple comparisons.

Results

Among the 7567 exams available for analysis, training level was missing for 284, 10 had no specialty identified, and 41 were before 2008. Therefore, a total of 7232 FLS tests were analyzed. Participants were 64% male, 6.4% junior residents (PGY 1-2), 84% senior residents (PGY3-5), 2.8% surgical fellows (PGY6), and 6.7% attending surgeons (Appendix 3.2.1). Specialties included 93% general surgery (GS), 6.2% gynecology, and 0.9% urology.

The Pearson correlation between cognitive and manual exam scores was 0.09. For the cognitive exam, the scores did not increase consistently each year, however, there was a significant increase in scores over the years (2008-2011 versus 2012-2016), and the most junior residents scored lowest. Participants with a GS background scored higher

than other specialties. For the manual skills, there was also no clear and consistent pattern in score differences over the years, however, there was a marginal decrease in scores over the years (2008-2011 versus 2012-2016), and junior residents scored higher. GS scored slightly higher than gynecology, but not urology (Table 1).

When looking at each task separately, different patterns emerged (Table 1). In the peg transfer task, the 2012-2016 group and junior residents had significantly higher scores than the 2008-2011 group and more senior trainees and attending surgeons. For pattern cutting and extracorporeal knot tying, the 2008-2011 group and junior residents had significantly higher scores than the 2012-2016 group and more senior trainees and attending surgeons. For the endoloop task, there were no significant differences in performance between the earlier and more recent cohorts or among training levels. For intracorporeal knot tying, there were no significant differences between the earlier and more recent cohorts, and junior residents scored highest. For peg transfer, there were no differences between specialties. For pattern cutting, endoloop, and intracorporeal knot-tying, GS scored slightly higher than Gynecology, but not Urology. For extracorporeal knot-tying, GS scored slightly higher than Urology, but not Gynecology.

When Pearson correlations between the tasks were calculated, all of them had moderate correlation. The lowest correlation was between endoloop and intracorporeal knot tying (0.32), and the highest correlation was between intracorporeal and extracorporeal knot tying (0.46).

Failure rates and patterns by level of training are shown in Table 2. 256 test-takers failed the exam and of those, 96 did a re-assessment (91 only cognitive, 4 only manual, and 1 both). The odds ratio of a PGY3+ passing was 1.8 (confidence interval

1.2-2.8) times higher than a PGY1-2. The internal consistency between tasks on the manual skills exam was 0.73. If any one of the tasks was removed, the Cronbach's alpha dropped to between 0.65 -0.71, depending on the task being removed.

Discussion

This study reports the results of FLS testing over the past 9 years in order to identify trends over time and to periodically monitor the exam and update it as needed. More experienced trainees continue to have a higher likelihood of passing the exam, which may be useful information for Program directors and trainees in planning when to take the FLS exam. We cannot tell from this analysis, however, if the content is still current and representative of the knowledge, skills, and level of performance expected. This would require an update of the content by reviewing all of the knowledge and skills that make up the fundamentals of laparoscopic surgery. In addition, a review of what defines the minimally qualified candidate in basic laparoscopy would also be needed to determine the pass/fail set point. This type of process was recently performed for the cognitive portion of the exam, and is in progress for the manual skills.

In our analysis, not surprisingly, higher PGY levels and attendings scored better than lower PGY levels on the cognitive part of the exam. In contrast, PGY 1-2 residents had better scores than senior residents and attendings on the manual skills portion of the test. This could be due to the fact that junior residents may have been more motivated to practice the skills and to use the established proficiency metrics as targets. This logic may also apply to fellows who feel the pressure to have excellent skills as they are expected to perform in the operating room and to teach residents. Furthermore, senior residents and attending surgeons tend to have more confidence in their skills if they perform

laparoscopic surgery regularly, and may not have practiced as much prior to taking the test. The actual difference in scores is quite small, and this is likely not to have any real significance from a clinical or test-taking perspective. More senior level trainees are still more likely to pass the FLS test, and this finding does not reflect a lack of validity since we do not have any data about the practice patterns of either of these groups in this cohort. This difference could also explain in part the Pearson's correlation of 0.09 between the cognitive and manual portions of the exam. The low correlation also indicates that the cognitive and manual components of FLS assess different domains of knowledge and skill for laparoscopy.

The internal consistency of the manual skills tasks is very good at 0.73, suggesting that the tasks are all measuring the same overall construct, in this case, basic laparoscopic technical skills. Removal of any one of the tasks, however, decreased the internal consistency, indicating that each task contributes something unique to the overall score. Also, each component of the manual skills exam measures a slightly different aspect of laparoscopic skills and there is no evidence to justify removal of any one task. The results indicate that the exam performs well and that it is not redundant.

Test takers with a General surgery background scored higher in manual exam than those with training in Gynaecology, but not Urology. Even though the difference in scores was statistically significant between GS and Gynecology, the actual difference was marginal and likely not significant in practice. For the cognitive exam, however, GS scored higher than other specialties, with about a 14% difference in the scores. Although the exam has always aimed to be non-specialty specific and to focus on fundamentals that would be common to all specialties, there may be an inherent GS bias in the questions.

The cognitive exam is currently being updated, with special attention to removing any questions that may have specialty-specific components.

The main limitation of this study is that the data analyzed did not include all of the FLS exams that were taken during that time period. We had data for 60% of test-takers, therefore, the ratios of the number of test-takers in different years and PGY levels could vary. This is due to issues in collecting and extracting the data using different systems and methods over the years. Based on a limited review of the various ways the data were incomplete, we don't have any reason to believe that the analysis presented in this paper has any specific bias, however, it is not possible to know this with absolute certainty. There was no pattern to the missing data and the sample is thought to be representative of the cohort of test takers during the time period studied.

This analysis was performed to ensure that the FLS exam continues to perform as well as it did in the past and to identify, in particular for the manual skills, any tasks that do not add to the overall exam. As in every profession, however, knowledge and skills evolve and the performance bar is often readjusted over time. In order to maintain the high quality of the exam and to apply the same rigor that has always been the sine qua non of the FLS program, the FLS Committee is also embarking on a review of the content of both the cognitive and manual skills portion of the exam in addition to a re-evaluation of the minimum performance level expected.

Acknowledgments

We would like to thank Carla Bryant, Christelle Menetrier and all of the SAGES FLS staff for their help in conducting this study.

Table 1: Trends over year and training level

	Mean Estimate*	Confidence Interval
Cognitive		
Year (2012-16 vs 2008-11)	66.7	62.4-71
PGY level (3-7 vs 1-2)	56.7	48.2-65.2
Specialty (GN vs GS)	-72.7	-81.5-(-63.9)
Specialty (UR vs GS)	-95.2	-117.7-(-72.8)
Manual		
Year (2012-16 vs 2008-11)	-6	-10.4-(-1.5)
PGY level (3-7 vs 1-2)	-19.4	-28.2-(-10.6)
Specialty (GN vs GS)	-12.4	-21.6-(-3.2)
Specialty (UR vs GS)	-11.2	-34.6-12.1
Peg transfer		
Year (2012-16 vs 2008-11)	7.1	5.8-8.4
PGY level (3-7 vs 1-2)	-9.5	-12.1-(-6.9)
Specialty (GN vs GS)	-1.8	-4.4-0.9
Specialty (UR vs GS)	-5.9	-12.7-0.9
Pattern cutting		
Year (2012-16 vs 2008-11)	-8.9	-11.4-(-6.4)
PGY level (3-7 vs 1-2)	-11.8	-16.7-(-7)
Specialty (GN vs GS)	-5.2	-10.2-(-0.17)
Specialty (UR vs GS)	7.6	-5.1-20.4
Endoloop		
Year (2012-16 vs 2008-11)	-0.6	-1.9-0.75
PGY level (3-7 vs 1-2)	2.3	-0.33-5
Specialty (GN vs GS)	-4.4	-7.2-(-1.7)
Specialty (UR vs GS)	-0.75	-7.8-6.3
Extracorporeal knot-tying		
Year (2012-16 vs 2008-11)	-8	-10.6-(-5.4)
PGY level (3-7 vs 1-2)	-10.75	-15.9-(-5.6)
Specialty (GN vs GS)	0.2	-5.1-5.5
Specialty (UR vs GS)	-16.2	-29.7-(-2.7)
Intracorporeal knot-tying		
Year (2012-16 vs 2008-11)	-2.4	-5.5-0.8
PGY level (3-7 vs 1-2)	-9.1	-15.4-(-2.9)
Specialty (GN vs GS)	-10.6	-17.1-(-4.2)
Specialty (UR vs GS)	-9.9	-26.4-6.5

PGY post-graduate year, PGY 6: Fellow, PGY7: Attending
GS General surgery, GN Gynaecology, UR Urology
*The differences in the scores between the two groups for Year and PGY level. The difference is significant if the confidence interval does not include 0.

Table 2: Failure rates and patterns by level of training

PGY Level	N	Failure-rate (%)	Pass cognitive/fail manual (%)	Pass manual/fail cognitive (%)	Fail both (%)
1	106	11(11.38)	0	11 (10.38)	0
2	358	18 (5.03)	2 (0.56)	16 (4.47)	0
3	954	30 (3.14)	5 (0.52)	25 (2.62)	0
4	2079	59 (2.84)	16 (0.77)	43 (2.07)	0
5	3043	71 (2.33)	12 (0.39)	57 (1.87)	2 (0.07)
6	207	14 (6.8)	2 (0.97)	12 (5.8)	0
7	485	66 (13.6)	23 (4.47)	40 (8.25)	3 (0.62)
PGY post-graduate year, PGY 6: Fellow, PGY7: Attending					

3.3 Multicenter Proficiency Benchmarks for Advanced Laparoscopic Suturing Tasks

Running Head: Proficiency Benchmarks for Suturing Tasks

Authors:
Elif Bilgic[1] BSc, Yusuke Watanabe*[1,2] MD, PhD, Dmitry Nepomnayshy[3] MD, Aimee Gardner[4] PhD, Shimae Fitzgibbons[5] MD, MEd, Iman Ghaderi[6] MD, Adnan Alseidi[7], MD, EdM, Dimitrios Stefanidis[8] MD, PhD, John Paige[9] MD, Neal Seymour[10] MD, PhD, Katherine M. McKendy[1] MD, MEd, Richard Birkett[3] MD, James Whitledge[11] BA, Erica Kane[10] MD, Nicholas E. Anton[8] MS, Melina C. Vassiliou[1] MD, MEd : for the Simulation Committee of the Association for Surgical Education

Institutions and Affiliations:
1. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre
2. Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine
3. Department of General Surgery, Lahey Hospital and Medical Center
4. University of Texas Southwestern Medical Center
5. MedStar Georgetown University Hospital
6. University of Arizona
7. Virginia Mason Medical Center
8. Carolinas Health Care System
9. LSU Health New Orleans Health Sciences Center
10. Baystate Medical Center
11. Tufts University School of Medicine

*Co-primary author

Conflicts of Interest and Sources of Funding:
All authors have no relevant conflicts of interests to disclose. The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Medtronic Canada. There was no funding or financial support for this work.

Corresponding author
Dmitry Nepomnayshy, MD
Department of General Surgery, Lahey Hospital and Medical Center
41 Mall Road, Burlington, MA 01805, USA

Abstract:

Background:
Advanced laparoscopic suturing(LS) tasks were developed based on a needs assessment. Initial validity evidence has been shown. The purpose of this multicenter study was to determine expert proficiency benchmarks for these tasks.

Methods:
6 tasks were included: needle handling(NH), offset-camera forehand suturing(OF), offset-camera backhand suturing(OB), confined space suturing(CF), suturing under tension(UT), and continuous suturing(CS). Minimally invasive surgeons experienced in LS completed the tasks twice. Mean time and median accuracy scores were used to establish the benchmarks.

Results:
Seventeen MIS surgeons enrolled, from 7 academic centers. Mean(95% CI) time in seconds to complete each task was: NH 169(149-189), OF 158(134-181), OB 189(154-224), CF 181(156-205), UT 379(334-423), and CS 416(354-477). Very few errors in accuracy were made by experts in each of the tasks.

Conclusions:
Time- and accuracy-based proficiency benchmarks for 6 advanced LS tasks were established. These benchmarks will be included in an advanced laparoscopic surgery curriculum currently under development.

Keywords: Proficiency, Competency, Performance, Assessment, Simulation, Suturing, and Laparoscopy

Background

Over the past two decades, laparoscopic surgery has advanced in both the range of procedures commonly performed and the skills required to perform these procedures safely. Despite these skills demands, many advanced laparoscopic procedures have become routine practice in general surgery. At the conclusion of training, however, the majority of general surgery residents may not be either comfortable or adequately prepared to perform advanced laparoscopic procedures. Most residents pursue fellowships, in many cases specifically to hone technical skills relevant to advanced laparoscopic surgery[104-106].

Although simulation has become an important part of residency training for a number of different procedures and skills, there are many aspects of surgical training for which few, if any, simulation opportunities exist. A comprehensive needs assessment identified a considerable gap between the advanced skills required to perform safe laparoscopic suturing in the operating room and the basic skills targeted by current simulators[107]. A recent survey of Fellowship Council program directors for non-ACGME accredited Minimally Invasive Surgery (MIS) fellowships in Canada and the US reported that 60% of program directors thought that graduates were not proficient in laparoscopic suturing at the beginning of their MIS fellowships[21].

Laparoscopic suturing is an essential technique required for a wide range of advanced laparoscopic procedures including closure of hiatal defects, peritoneal and mesenteric closure, and bowel anastomosis[108]. It is also needed in order to manage a variety of intraoperative complications. This skill, however, is one of the most challenging for surgical trainees to master. Reasons for this may include variability of

suturing techniques, lack of advanced models and training curricula that reflect clinical complexities, and limited trainee exposure to advanced MIS procedures during residency[37, 38].

Several advanced laparoscopic suturing tasks have been developed in the past including synthetic bowel models, virtual reality simulations with motion-based metrics, explanted porcine small intestine models, and live animal models[43, 108, 109]. These models have several limitations in terms of cost, practical usability, and lack of robust validity evidence to support them as measures of advanced laparoscopic suturing skills. Based on these gaps, and a previous needs assessment, advanced laparoscopic suturing (ALS) skills were identified that are not effectively taught and assessed using current bench-top modules[23]. Subsequently, 6 ALS tasks were developed using inexpensive, readily available materials: needle handling (NH), offset-camera forehand suture (OF), offset-camera backhand suture (OB), confined space suture (CF), suturing under tension (UT), and continuous suturing (CS). Performance metrics of these tasks were shown to have preliminary validity evidence as measures of advanced suturing skills[107, 110].

Proficiency-based training provides an optimal approach for technical skills acquisition by enabling goal-directed practice to pre-determined levels of expertise, leading to uniform skill acquisition by trainees regardless of individual learning curves[100, 111, 112]. The purpose of this multicenter study was to determine expert-derived performance benchmarks for the ALS tasks in order to design a proficiency-based advanced laparoscopic surgery curriculum.

Methods

This was a multi-institutional prospective study conducted at health care institutions in the United States and Canada and approval from the Institutional Review Boards at all sites was obtained. Experienced laparoscopic surgeons on MIS or Bariatric services at all sites were recruited to perform the ALS tasks. Participants were considered eligible to participate if they performed 25 or more laparoscopic suturing cases per year as responsible faculty surgeon, without the use of a robot or other assist device (such as the endostichTM).

After consent for participation was obtained, all surgeons performed 2 consecutive repetitions of the 6 advanced laparoscopic suturing tasks placed in the FLS trainer box (Limbs & Things, Savannah, GA). All participants performed the Fundamentals of Laparoscopic Surgery (FLS) intra or extra-corporeal suturing task (surgeons' choice) for 1-3 minutes before testing as a warm up prior to proceeding with ALS tasks. They viewed instructional videos explaining each of the ALS tasks before performing them. After completion of the tasks, participants were asked to complete a questionnaire that solicited demographics, and clinical experience information. The tasks were timed by two raters at each institution, and accuracy was assessed by two raters (Y.W and E.B) at host institution. The inter- rater reliabilities were calculated using a two-way random effects model of intraclass correlation coefficients (ICCs).

Advanced Laparoscopic Suturing tasks

The 6 tasks were developed based on a needs assessment study that demonstrated the necessity for simulated advanced suturing tasks. The tasks included needle handling (NH), offset-camera forehand suture (OF), offset-camera backhand suture (OB), confined space suture (CF), suturing under tension (UT), and continuous suturing (CS; Table

1)[107, 110]. In NH, participants pass a needle through six holes of a circle starting from the top right, in a sequential, counter-clockwise fashion. In OF and OB, participants perform forehand and backhand suturing respectively, first a double throw and then two single throws with the camera offset from the standard view. In CF, participants perform forehand suturing in a confined space. In UT, participants perform 3 interrupted sutures to close a wide defect, while they decide on suture length and knot type; they have to have 3 ties for each suture. In CS, participants perform suturing to close a defect in a continuous fashion. These tasks are inexpensive and made using readily available materials. More details on the development of these tasks have been published previously[107, 110].

The tasks are assessed using time and accuracy. Time represents the time taken to complete the tasks and accuracy is measured using predefined penalties for each task. The accuracy scores for each task include: the number of times the needle is dropped outside the field of view (NH); gaps in the closure, the distance of the suture from the pre-marked dots and where the suture is placed, and the security of the knots (OF, OB, CF, UT, CS); and number of skipped dots (CS).

Development of proficiency benchmarks

The proficiency benchmarks for each task were determined based on the performance of surgeons using time and accuracy (penalty) metrics. Mean performance times and median accuracy scores for all subjects were used. For time, any outliers beyond 2 standard deviations from the mean were excluded. The trimmed mean time was then used as the basis for determining the recommended benchmarks, along with the median accuracy scores.

Results

A total of 17 surgeons participated in this study from 7 academic centers in North America. Seventy-six percent were male and 76% had completed an MIS or Bariatric surgery fellowship. The median number of years in practice was 6 (2-15), and 58% reported that they perform over 101 cases per year using intracorporeal suturing without an assist device (Table 2). All of the participants used intracorporeal suturing.

The inter-rater reliability was 0.99. The mean (95% CI) time in seconds to complete each task was: NH 169 (149-189), OF 158 (134-181), OB 189 (154-224), CF 181 (156-205), UT 379 (334-423), and CS 416 (354-477). For the CS task, all of the participants used an intracorporeal suturing technique. For the median accuracy scores; there were no needles dropped outside the field of view for NH, no gaps between sutures for all tasks, knot was secure for all tasks, the suture was 1mm-off from the pre-marked dots for OF, OB and CS, no sutures were off the pre-marked dots for UT, and there were no skipped dots for CS. The performance-based proficiency benchmarks for each task for both time and accuracy are shown in Table 3.

Discussion

This multicenter study established time- and accuracy benchmarks for each of the 6 advanced laparoscopic suturing tasks based on the performance of experienced minimally invasive surgeons. Since this was a multicenter study, recruitment of the surgeons was feasible, and the results are more generalizable across North America. Also, a multicenter study allows for the evaluation of performance without taking into account differences in practice patterns within individual training programs[113]. The Fundamentals of Laparoscopic Surgery (FLS) is a standardized program to develop and

assess basic laparoscopic skills. Evidence supports the use of low-fidelity bench-top simulators to develop basic fundamental laparoscopic skills and transfer of these skills to the clinical environment has been repeatedly demonstrated[13, 18-20, 99, 114]. However, from a needs assessment that was conducted, it was found that there is a need for more advanced tasks that could teach and assess more advanced skills that are required for clinical practice that are not currently available with the current simulation models, including suturing[23]. Therefore, 6 cost-effective advanced laparoscopic suturing tasks that better reflect complexities in practice were developed. As the process of developing an advanced laparoscopic surgery curriculum is underway, the developed benchmarks will be included in the psychomotor skills portion of the curriculum. With these benchmarks, trainees will be able to train with specific goals as to what time and accuracy they should achieve in order to be considered proficient in the tasks.

Proficiency-based training has been shown to increase learner motivation, along with attendance to the skills laboratory, and leads to improved performance when compared to training that does not clearly establish objectives. Stefanidis et al. has demonstrated this phenomenon for basic inanimate and virtual reality simulation tasks. In their study, they found that with performance goals, residents were more motivated to participate in a simulation curriculum, which led to improvements in their performance[115]. Similar results were seen in a study by Madan et al, where they compared laparoscopic training that is goal-directed versus without goals using basic laparoscopic tasks. The goal directed group achieved significantly higher scores on a post-test than the group that trained without performance goals[116].

FLS has established time- and accuracy-based proficiency benchmarks to allow for a goal-oriented training of the surgical trainees[92]. Stoller et al has done a study to evaluate the effectiveness of the FLS proficiency-benchmarks compared to another training method which provided learning goals, and they found that both methods of teaching had significant improvements in performance, suggesting that different methods could be used to train using the FLS tasks[117]. Sroka et al. compared FLS and intra-operative laparoscopic assessment scores of residents who underwent training using the FLS proficiency-based curriculum versus no training. They found that the training group had significant improvements in their score for both FLS and intra-operative skills[18].

The FLS programs' time and accuracy metrics were found to have excellent validity in terms of differentiating different levels of residents, and consistency of their scores[91, 101, 118]. However, within the proficiency-based curriculum that is being developed, for the purposes of formative feedback, other metrics could be incorporated to facilitate skill acquisition. Stefanidis et al provided evidence for incorporation of secondary task metrics, where residents trained beyond the proficiency benchmarks, to accomplish expert levels on visual spatial tasks (automaticity). They compared FLS proficiency-benchmarks alone versus with automaticity measures and found that participants had better scores in the operating room when their training included automaticity. This study showed that going beyond proficiency-benchmarks, and adding a different metric that went beyond the initial proficiency allowed for better skill acquisition[119]. Scoring of the ALS tasks is based on time and accuracy; however, other metrics could also be included for training of surgical trainees.

The proficiency benchmarks were developed using performance time and maximum error score that is allowable. It has been shown that using this method over using a calculated score enables a more feasible approach to determining current performance level and the improvement necessary to achieve proficiency[92]. This allows for a real time scoring during the training, and immediate feedback. We chose 2 standard deviations from the mean as our cutoff, since this allows for a better estimation of the true mean performance times, which is used by most researchers. It has been shown that surgeons that score very high or very low were included, this would not be representative of the proficient surgeon population[120].

In conclusion, the ALS suturing tasks address the perceived gap in training to better prepare residents for clinical practice. In this study, we established proficiency benchmarks for these tasks which will be incorporated into an advanced laparoscopic surgery curriculum. Future studies will assess and hopefully build the validity of this curriculum.

Table 1: Advanced laparoscopic suturing tasks

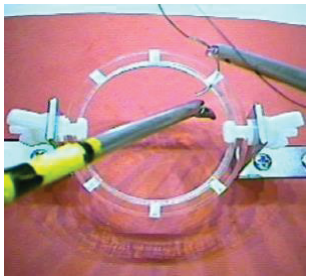
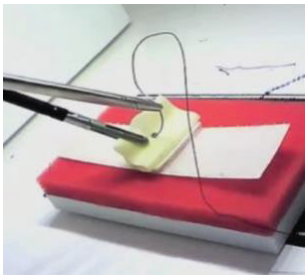
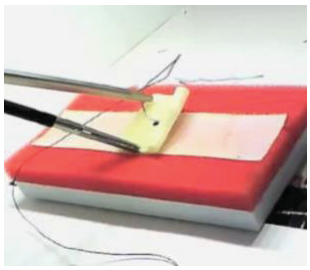
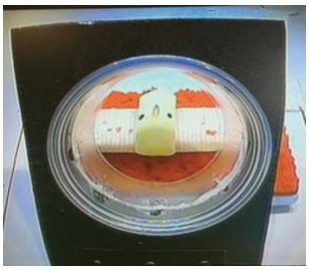
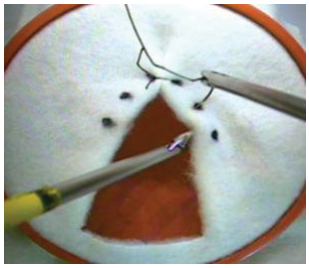
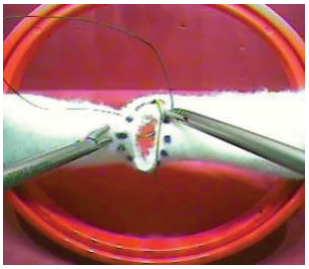
Needle Handling: Manipulating the needle properly for desired angles and directions 	**Off-set Forehand Suturing:** Forehand suturing with an off-set camera position 
Off-set Backhand Suturing: Backhand suturing with an off-set camera position 	**Confined Space Suturing:** Suturing within a confined space 
Suturing Under Tension: Suturing a tissue under tension (e.g. simulates Nissen fundoplication) 	**Continuous Suturing:** Suturing as a continuous closure (e.g. simulates bowel anastomosis) 

Table 2: Characteristics of 17 surgeons. Results are presented as n (%) or median [interquartile range].

	N (%)
Male/ Female	13 (76) / 4 (24)
Years in Practice	6 [2-15]
MIS/Bariatric Fellowship	
Yes/ No*	13 (76) / 4 (24)
Lap suturing experience (annual)	
26 -50 cases	3 (18)
51 -100 cases	4 (24)
101+ cases	10 (58)

* Surgeons were experienced 14, 23, 24, and 27 years in practice, respectively.

Table 3: Performance-based proficiency benchmarks based on mean time and median accuracy scores of MIS/Bariatirc surgeons' performances

Task	Mean time (95% CI)*	Median accuracy scores (25th-75th percentile)
Needle handling	169 (149-189)	0 (0-0) needles dropped outside field of view
Offset-camera forehand suture	158 (134-181)	1 (0-2) mm off from the dots 0 (0-0) mm gaps in closure 0 (0-0) knot security error
Offset-camera backhand suture	189 (154-224)	
Confined space suture	181 (156-205)	1 (0-2) mm off from the dots 0 (0-0) mm gaps in closure 0 (0-1) knot security error
Suturing under tension	379 (334-423)	0 (0-0) mm off from the dots 0 (0-0) mm gaps in closure 0 (0-0) knot security error
Continuous suturing	416 (354-477)	1 (0-3) mm off from the dots 0 (0-0) mm gaps in closure 0 (0-0) knot security error 0 (0-0) skipped dots

*Time reported in seconds

3.4 Development of a Model for the Acquisition and Assessment of Advanced Laparoscopic Suturing Skills using an Automated Device

Running Head: Laparoscopic Suturing Task for Device-assisted Suturing

Authors:
Elif Bilgic[1] BSc, Madoka Takao[1,2] MD, Pepa Kaneva[1]MSc, Satoshi Endo[1,3] MD, Toshitatsu Takao[1,2] MD, Yusuke Watanabe[4] MD, Katherine M. McKendy[1] MD, Liane S. Feldman[1] MD, Melina C. Vassiliou[1] MD

Institutions and Affiliations:
1. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, Quebec, Canada
2. Department of Gastroenterology, Kobe University Graduate School of Medicine, Kobe-shi, Hyogo, Japan
3. Department of Frontier Surgery, Chiba University Graduate School of Medicine, Chiba-shi, Chiba, Japan
4. Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan

Conflicts of Interest and Sources of Funding:
The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Medtronic Canada. Dr Melina C. Vassiliou received Investigator Sponsored Research in the form of Endo Stitch™ devices and sutures from Medtronic. Elif Bilgic, Pepa Kaneva, and Drs Madoka Takao, Toshitatsu Takao, Yusuke Watanabe, Katherine M. McKendy, Satoshi Endo, and Liane S. Feldman have no relevant conflicts of interests to disclose.

Corresponding author
Melina C. Vassiliou, MD, MEd
McGill University Health Centre
1650, Cedar Avenue, L9. 313, Montreal, QC H3G 1A4, Canada.
TEL: +1-514-934-1934, ext. 44330, FAX: +1-514-934-8210
E-mail: melina.vassiliou@mcgill.ca

Abstract:

Background:
Needs assessment identified a gap regarding laparoscopic suturing(LS) skills targeted in simulation. This study collected validity evidence for an advanced LS task using an Endo Stitch™ device.

Methods:
Experienced(ES) and novice surgeons(NS) performed continuous suturing after watching an instructional video. Scores were based on time and accuracy, and Global Operative Assessment of Laparoscopic Surgery(GOALS). Data are shown as Medians[25-75th percentile](ES vs NS). Inter-rater reliability was calculated using intra-class correlation coefficients(confidence interval).

Results:
Seventeen subjects were enrolled. ES had significantly greater task(980[964-999] vs 666[391-711], p-value 0.0035) and GOALS scores(25[24-25] vs 14[12-17], p-value 0.0029). Inter-rater reliability for time and accuracy were 1.0 and 0.9(0.74-0.96) respectively. All experienced surgeons agreed that the task was relevant to practice.

Conclusion:
This study provides validity evidence for the task as a measure of LS skill using an automated suturing device. It could help trainees acquire the skills they need to better prepare for clinical learning.

Key words: laparoscopic suturing; surgical training; competency based education

Background

Suturing is a skill that is required to perform a variety of different laparoscopic procedures across many different surgical specialties. Surgical trainees, however, find that laparoscopic suturing (LS) is a difficult skill to acquire and master. Simulation is often used to help trainees practice difficult skills, or to transfer part of the learning curve for these skills outside of the operating-room, but there is a lack of simulation platforms for this advanced skill. The Fundamentals of Laparoscopic Surgery program provides a platform to teach and assess basic laparoscopic skills, including basic laparoscopic suturing training[92, 98, 121]. However, in the operating-room, the conditions are more variable and nuanced and the skills needed for suturing are more complex and advanced.

A previous needs assessment identified a gap between the laparoscopic suturing skills needed in the operating-room compared to what is taught and assessed in the various simulation platforms[21, 23]. Based on this, a set of advanced LS models using free needles was developed and tested for validity evidence[107, 110, 122]. The models, however, were not adapted to be used with automated suturing devices such as Endo Stitch™. The needs assessment also revealed a gap with regards to learning how to use suturing devices, which are also commonly used in the operating-room[22]. Simulation platforms that are currently available for this are limited in terms of number, cost and validity evidence[123, 124]. Therefore, there is a need to develop cost-effective tasks and metrics that could be used to teach and assess the skills required for device-assisted laparoscopic suturing skills.

The purpose of this study was to design and collect validity evidence for an advanced LS task and metrics using the Endo Stitch™ device and to develop an assessment tool that could be used for assessment and feedback of Endo Stitch™ skills.

Methods

Task Description:

The task, based on our previous advanced LS models for free needle, was developed with readily available and low cost materials, and in consultation with experts. At first, we tried the fabric that was used for the free needle tasks; however, due to the differences in the properties of the free- needle compared to the needle of the device, the Endo Stitch™ needle would get stuck and the needle would come off of the device. Therefore, in order to develop the task, various materials were tested such as different types of fabrics and sponges, and the Endo Stitch™ only worked with one type of sponge. Therefore, the task was developed with the sponge, and a gap of seven cm in the middle was created with five pre-marked black dots on each side of the gap (1cm apart from one another) (Figure 1). The task was placed in the Fundamentals of Laparoscopic Surgery trainer box (Limbs & Things, Savannah, GA). The participants were asked to start from the top and perform continuous suturing with intracorporeal knot-tying at the beginning and at the end; they had to go through the pre-marked black dots.

Study Design:

Experienced (ES) and novice surgeons (NS) performed the task once after watching an instructional video (the video can be made available to the readers upon request). At first, during the data collection phase, determination of participant experience was based on the number of cases they have done using the Endo Stitch™. NS were allowed to warm up for up to three minutes to get use to the Endo Stitch™

device. ES completed the task a second time, in order to establish proficiency benchmarks. The participants also completed a questionnaire on demographics and perceived educational value. Scores were based on time and accuracy and assigned by direct observation. The cutoff time was 1200 seconds. Accuracy scores were based on knot security, gap in the incision, distance from the pre- marked dots, and skipped dots. Participant performances were also recorded and assessed by one rater using the Global Operative Assessment of Laparoscopic Surgery (GOALS) instrument[125]. For analysis, participants who received a GOALS score of ≥20 were considered as experienced surgeons.

Validity evidence was gathered by comparing task scores and GOALS scores between the two groups. Data are shown as Medians [25-75th percentile] (ES vs NS). Inter-rater reliability was calculated using intra-class correlation coefficients (confidence interval); two raters assessed accuracy for all of the completed tasks and assessed time using the recorded performances of five participants. Comparison of task scores, accuracy scores alone, and GOALS scores between ES and NS were made using Mann–Whitney U test. Pearsons correlation was used to correlate task scores to GOALS scores. The proficiency benchmarks were determined by using the 75^{th} percentile of the ES performances, using the score from their second task completion.

For the creation of the 'Endo Stitch™ assessment tool', experienced surgeons who completed the task were interviewed in order to understand the skills that are important when performing LS using Endo Stitch™. ES were asked to comment on the key steps required to perform LS with Endo Stitch™ in this task and watch a video of a novice performing the task while commenting on how the person could improve their

performance. After the creation of the tool, two raters assessed the video-recorded performances to assess the inter-rater reliability. Comparison of 'Endo Stitch™ assessment tool' scores between ES and NS were made using Mann–Whitney U test. Pearson correlation was used to correlate 'Endo Stitch™ assessment tool' scores to task scores and GOALS scores. All analysis was done using SAS 9.3 and SPSS V20.0.0. A p value ≤ 0.05 was considered statistically significant.

Power analysis:

Pilot performance data for continuous suturing without using an automated suturing device from 18 senior surgical residents and 13 MIS trained surgeons showed significant differences in performance in both training platforms[110]. In order to show a similar difference between the two groups in performance with an α of 0.05 and a power of 80%, with 2-sided testing, a total of 8 participants per group is required.

Results

Seventeen subjects (9 ES, 8 NS; median age 34, 76% male) were enrolled. All of the ES completed an MIS/Bariatric fellowship. Table 1 shows comparisons between ES and NS performing the new Endo Stitch™ task. Compared to NS, ES had significantly greater scores (980[964-999] vs 666[391-711], p-value 0.0035). ES made errors in 'gaps in incision' and 'distance from pre-marked dots', however, NS made significantly more errors overall (6[4-9] vs 20[12-23], p-value 0.014). ES also received significantly better scores in GOALS (25[24-25] vs 14[12-17], p-value 0.0029). The Pearson correlation between task scores and GOALS scores was 0.89. Inter-rater reliability for time and accuracy were 1.0 and 0.9 (0.74-0.96) respectively. All experienced surgeons agreed or strongly agreed that the task was relevant to practice, and that it could be used to improve

device-assisted suturing skills. Two of the nine experienced surgeons only had one trial, therefore the performances of the other seven participants were used to determine the benchmarks. The proficiency benchmarks can be seen in Table 2.

From interviewing seven ES, an assessment tool with 13 items was developed, with a rating scale of 0-2 (0-does poorly, 1-needs some improvement, 2-does well) and a total score of 26. The tool structure was based on assessment tools that were developed for the free-needle advanced LS tasks[126]. It includes generic and specific items regarding steps needed when suturing using an Endo Stitch (Table 3). Compared to NS, ES had significantly greater 'Endo Stitch™ assessment tool' scores (26[24-26] vs 16[11-20], p-value 0.0032). The Pearson correlation between 'Endo Stitch™ assessment tool' scores and the task scores and GOALS scores was 0.95 and 0.94 respectively. The inter-rater reliability was 0.9(0.61-0.97).

Discussion

FLS was developed to address basic laparoscopic skills, including suturing. However, there was a need for advanced platforms to teach and assess LS skills, and in particular device-assisted LS. In order to address this need, and to model the skills needed in the operating-room, this new task was developed. This study provides preliminary validity evidence for an advanced laparoscopic suturing task using Endo Stitch™; the inter-rater reliability for the metrics are high, the task is able to differentiate experts from novices, and the task metrics are highly correlated with the video-taped assessment of their task performance using GOALS. The experienced surgeons agreed on the relevancy of the task to practice, and its potential to improve device-assisted suturing skills. Also,

an assessment tool that assesses the Endo Stitch™ suturing skills in the task was developed, with good preliminary validity evidence.

There are some simulation platforms available for LS. Specifically, for usage with automated devices, the available platforms are a lot less in number, costly, and have limited validity evidence.[123] Our task was developed from readily available, low-cost materials, and it can be used in a variety of bench-top boxes. If the purpose of using the task is for teaching, individuals can tailor it to their needs. Also, since there are different types of automated suturing devices apart from the Endo Stitch, this task could be used with the various devices and validity evidence could be demonstrated.

One of the metrics that we used was GOALS. Even though GOALS was originally developed for operative assessment and that it does not have prior validity evidence for assessing suturing skills in our context (device-assisted suturing in simulation), we were able to show that it has good correlation with the other metrics, and that it could differentiate between ES and NS. In addition, even though we used 1 rater for GOALS assessment, which prevented us from providing inter-rater reliability, we were still able to show some psychometric properties of the tool in our context.

During the data collection phase, we defined ES as individuals who perform at least 25 cases using the Endo Stitch™ device per year. For the data analysis, we used the GOALS scores of the individuals and defined ES as individuals who received a score of ≥20. We used the GOALS criteria to determine expertise because case numbers do not always provide an accurate representation of someones' skill level and that using an assessment tool could help us better define expertise. However, in our study, all surgeons who fit ES criteria for case numbers also received a GOALS score of ≥20.

The Endo Stitch™ assessment tool was developed based on expert opinions

regarding skills needed to perform LS in our task and validity evidence was demonstrated. Therefore, for the assessment of this task, either the tool and/or the time and accuracy scores could be used. However, this tool could also be used for formative assessment, in order to give feedback and improve the skill of the novices. In a previous study, using the previously developed advanced laparoscopic suturing tasks for free-needle, we compared skill improvement after training on the suturing models using assessment tools that we based our Endo Stitch™ assessment tool on (without a guide) versus guidance from an expert.[126, 127] It was found that both groups improved their suturing skills, suggesting that the items of the tools accurately reflected aspects of skill that experts considered to be important. Therefore, further investigations could be conducted in order to understand the value of our assessment tool for formative feedback.

This study has some limitations. First, the study was done at a single-institution, which could limit the generalizability of our results to other training programs. Second, we have not investigated how task performance correlates with operating-room performance and if the skills gained from this task transfer to the operating-room.

This study provides validity evidence for the task, its metrics, and the assessment tool as measures of LS skill using an automated suturing device. Incorporating this task into the training curricula could help trainees acquire the skills they need to be better prepared for and maximize clinical learning in the operating room.

Table 1: Comparison between Experienced and Novice surgeons

	Experienced	Novice	p-value*
Task score	980[964-999]	666[391-711]	0.0035
Accuracy	6[4-9]	20[12-23]	0.014
GOALS	25[24-25]	14[12-17]	0.0029

Experienced or Novice in performing suturing with Endo Stitch™
Total task scores; Accuracy count only
GOALS Global Operative Assessment of Laparoscopic Skills
Median[25-75th percentile]
*Significance when $p<0.05$

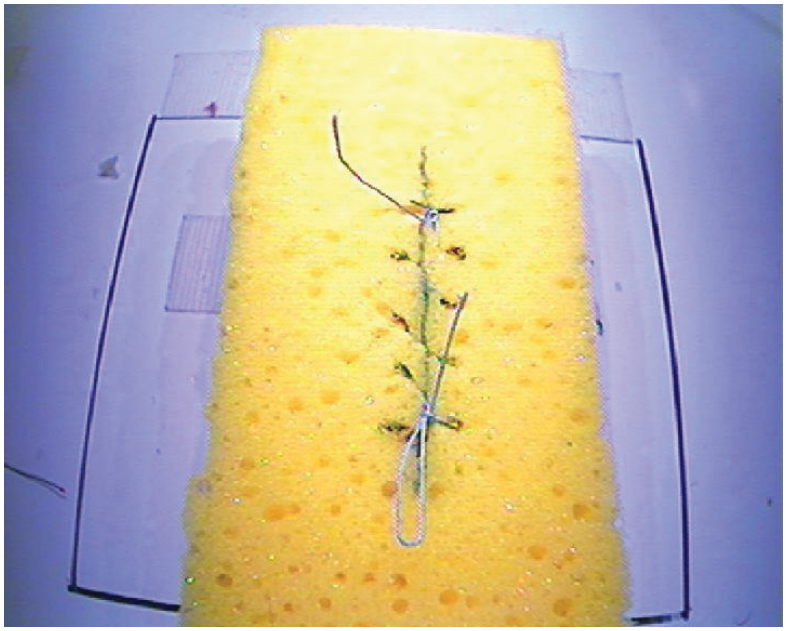
Table 2: Proficiency benchmarks determined through the performances of experienced surgeons (N=7)

Time	Accuracy
181	1 mm off from the dots 4 mm gaps in closure 0 knot security error 0 skipped dots

Table 3: The Endo Stitch™ assessment tool items for the task

1) Selects an appropriate suture length for the task
2) Knows how to use the Endo Stitch instrument (i.e. how to toggle the needle, how the mechanism works)
3) Anticipates the angle of the Endo Stitch when penetrating the target tissue for correct orientation of the needle
4) Understands the relationship between the suture and the Endo Stitch to tie a knot
5) Leaves suture tail an appropriate length when tying the knot
6) Uses the non-dominant hand as a post while tying the knot
7) Uses the dominant hand holding the Endo Stitch to tie around the post to make a knot
8) Ties the knots securely (tightly)
9) Coordinates the use of both hands
10) Takes the bites in the right orientation (i.e clockwise fashion to avoid leakage)
11) Keeps the excess suture out of the way while tying the knots and running the suture
12) Keeps the target tissue steady while suturing
13) Tightens the suture line and maintains appropriate tension while running the suture

Figure 1: The Endo Stitch™ task



CHAPTER 4: ASSESSMENT OF OPERATIVE PERFORMANCE OF LAPAROSCOPIC SUTURING

4.1 Pre-amble

In the previous chapters, we investigated various simulation platforms that could be used to assess both fundamental and advanced laparoscopic suturing (LS), for free-needle and device-assisted suturing. Simulation is a great environment for trainees to learn and practice as much as they need without time constraints or issues regarding patient safety. However, simulation training should have a benefitial impact on the operative-performance of the trainees. We also need to be able to assess trainees in the operating room (OR) accurately in order to track their progress, determine if their LS skills have improved, and to see if they are ready for the next step.

The assessment of LS skills in the OR comes with its own challenges. First, surgical educators have to determine which assessment tool fits their purpose of usage (feedback versus decision-making), and has evidence for the assessment conditions of their institution (do they have resources to have observers as raters or can only the attending surgeon of the case complete the assessment; is it feasible for them to do direct observation, or do they have resources to record videos for video-assessments etc). The first part of this chapter will address this by completing an in-depth review of the literature to identify and evaluate the validity evidence of the assessment tools developed or used to assess LS or procedures that require LS. This way, we will develop a guideline for surgical educators to help them in selecting an assessment tool that fits their purposes.

Second, it has been shown that for operative-assessments, 1 assessment is not enough to capture the true performance of the trainees. This is due to the fact that operative performance can be effected by factors such as the rater, case difficulty, procedure type etc, which are not related to the trainee performance, yet still effect the score they receive. Therefore, the second part of this chapter will focus on determining and minimizing the effect of raters and cases on the assessment scores, which will allow for an assessment that reflects the true operative-performance of trainees. This will be done by using an assessment tool called the Global Operative Assessment of Laparoscopic Skills (GOALS) that has evidence for assessment of procedures that require LS.

4.2 A Comprehensive Review of Assessment Tools for Laparoscopic Suturing

Running Head: Assessment Tools for Suturing

Authors:
Elif Bilgic[1] BSc, Satoshi Endo[1] MD, Ekaterina Lebedeva[2] MLIS, Madoka Takao[1] MD, Katherine M McKendy[1] MD, Yusuke Watanabe[3] MD, Liane S Feldman[1] MD, Melina C Vassiliou[1] MD

Institutions and Affiliations:
3. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada
4. The Henry K.M. De Kuyper Education Centre, McGill University Health Centre, Montreal, QC, Canada
5. Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan

Conflicts of Interest and Sources of Funding:
All authors have no relevant conflicts of interests to disclose. The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Medtronic Canada. There was no funding or financial support for this work.

Corresponding author:
Melina C. Vassiliou, MD, MEd
McGill University Health Centre
1650, Cedar Avenue, L9. 313, Montreal, QC H3G 1A4, Canada.
TEL: +1-514-934-1934, ext. 44330, FAX: +1-514-934-8210
E-mail: melina.vassiliou@mcgill.ca

Abstract:

Background:
A needs assessment identified a gap in teaching and assessment of laparoscopic suturing(LS) skills. The purpose of this review is to identify assessment tools that were used to assess LS skills, to evaluate validity evidence available, and provide guidance for selecting the right assessment tool for specific assessment conditions.

Methods:
Bibliographic databases were searched till April 2017. Full-text articles were included if they reported on assessment tools used in the operating-room(OR)/simulation to (1) assess procedures that require LS, or (2) specifically assess LS skills.

Results:
Forty-two tools were identified. 26 were used for assessing LS skills specifically, and 26 for procedures that require LS. Twenty-eight were global rating scales, 9 checklists, and 5 error rating scales. Tools had the most evidence in internal structure and relations to other variables, and least in consequences.

Conclusion:
Through identification and evaluation of assessment tools, the results of this review could be used as a guideline when implementing assessment tools into training programs.

Keywords: laparoscopic suturing; assessment; assessment tool; surgical training

Background

Laparoscopic suturing (LS) is as an advanced skill that is performed commonly during many different procedures across various surgical specialties. Yet, methods to explicitly teach and assess LS skills are limited [21, 23, 37, 38]. A needs assessment identified that there is a gap between what is targeted by current platforms versus what is needed in the clinical setting.[23] Several instruments have been developed to measure LS skills both in a simulated environment and in the operating room (OR), [43, 109, 128] however, the amount and quality of the evidence for their validity is highly variable and sometimes lacking all together.

There are different settings in which an assessment could be conducted: during the performance of simulated tasks on a bench-top, ex-vivo or in-vivo animal model, or in the OR [94, 107, 108, 110]. Depending on the consequences and intended uses of the assessment, validity evidence should be provided for each setting separately, and evidence from one setting cannot necessarily be generalized to another setting. The purpose of this review is to identify assessment tools that have been used to assess LS skills, or to assess procedures that require LS, and to evaluate the validity evidence available for the tools using the contemporary framework of validity [40, 129]. This information could be used by Program Directors and other surgical educators as a guide for selecting the appropriate assessment tool that meets the needs of specific assessment conditions.

Materials and Methods

Search strategy

Bibliographic databases MEDLINE, PubMed, Embase, PsychINFO, Scopus, ERIC via EBSCO, LILACS, The Cochrane Library, RDRB, and CINAHL were searched for full-text articles published between January 1990 and November 2016. The MEDLINE database was re-searched in April 2017 for any new studies that could be included. Search strategies were developed with the assistance of a health sciences librarian (E.L.). The search strategy included the use of key words and relevant indexing to identify articles discussing development and/or validation of assessment tools for simulation involving LS or that require laparoscopic suturing. The full MEDLINE strategy (Appendix 4.2.1) was applied to all databases, with search term and syntax modifications as needed. Citation searches were carried out to identify further studies.

Study selection

Full text articles, English or French, were considered. Studies were included if they reported on assessment tools (1) used in the OR/simulation to assess procedures that require LS, or (2) used in the OR/simulation to specifically assess LS skills. Reviews, abstracts, and editorials were excluded.

Data extraction

The reporting of the review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) standards. All studies were screened by two reviewers (E.B. and S.E.) in an independent fashion. Any disagreements were resolved by discussion and consensus. Full-text analysis was completed on articles that were included. Extracted information included study characteristics, characteristics of performance assessment tools, and 5 sources of validity evidence (content, response process, internal structure, relations to other variables, and consequences) [39]. The

validity evidence is rated according to the 'Criteria for Rating Validity Evidence by Ghaderi et al.[39] Each of the 5 sources receives a score from 0-3, for a total score of 15. The evidence is graded as; 1-5 limited evidence, 6-10 moderate evidence, and 11-15 strong evidence. The extracted information was reported in a similar way to Watanabe et al [130].

Results

Study characteristics

The primary search identified 3432 studies. After title and abstract screening, 397 underwent full-text review, of which 68 met our inclusion criteria (Figure 1)[3, 31, 32, 51, 68, 73, 131-192]. Study characteristics are listed in Table 1.

Assessment tool characteristics

Twenty-eight unique and 14 modified (from existing) assessment tools were identified, for a total of 42 tools. They included: 28 global rating scales, 9 checklists, and 5 error rating scales (Table 2). There were 15 generic, 20 specific, and 7 hybrid tools. Thirty-six were assessed using recorded videos, 14 by direct observation (7 observers, 9 attending surgeons, 6 self), and 8 were both. For the ones that were used to assess LS skills specifically, 3 were in the OR and 23 in a simulated environment. For the ones assessing procedures that require LS, 17 were in the OR and 9 in simulation. The types of simulation included box (inanimate tasks, ex-vivo, virtual and augmented reality), live porcine, and cadaver models. The types of intra-operative procedures most often assessed were laparoscopic hysterectomy, laparoscopic Roux-en Y gastric bypass, laparoscopic inguinal hernia repair, and laparoscopic Nissen fundoplication. Appendix 4.2.2 provides

information regarding the types of suturing and knot tying that were used as well as if the suturing was handsewn or device-assisted for each assessment tool.

For assessment of LS skills specifically, in the OR, a modified version of the Objective Structured Assessment of Technical Skills-Global rating scale (OSATS-GRS)[3], the original Global Operative Assessment of Laparoscopic Surgery (GOALS), and the original Van Sickle LS error rating tool were used. In simulation, the LS Checklist by Moorthy, OSATS-GRS, GOALS, and various modifications of these tools were used most commonly.

For assessment of procedures that require LS, in the OR, GOALS, OSATS-GRS, and modifications of these 2 tools, and a Generic Error Rating Tool (GERT) were used often. In simulation, OSATS-GRS, Global rating scale of operative skill (GRS- OS), and Assessment of Laparoscopic Roux-en Y gastric bypass (ALRYGB) were used commonly.

Validity evidence

The unitary framework of validity was used[39, 40, 130]. Summarization of validity evidence that is rated according to the 'Criteria for Rating Validity Evidence' by Ghaderi et al can be found Tables 3 and 4 for tools used in the OR, and Appendix 4.2.3 for tools used in simulation setting[39]. Appendix 4.2.4 summarizes the validity evidence with more detail.

Content

All of the studies provided a list of the items that were used to assess skill. For assessment of LS skills specifically, in the OR, none of the 3 tools were developed for operative LS assessment. In simulation, 'LS and Intracorporeal Knot Tying-Global

Rating Scale (LSIKT-GRS)', 'LS and Intracorporeal Knot Tying-Checklist (LSIKT CL)', 'Intracorporeal Knot Tying- Checklist (IKT-CL)', 'Laparoscopic suturing-Checklist (LS-CL)', and 'Assessment for suturing of vaginal cuff (ASVC) (Weizman 2014)' were developed to assess LS in a simulation task. All of them used expert judgment to develop the tools. IKT-CL, LS-CL, and ASVC also used a consensus method such as Delphi.

For assessment of procedures that require LS, in the OR, GOALS-Groin Hernia(GH), Transabdominal Preperitoneal procedure-Checklist (TAPP-CL), Bariatric Objective Structured Assessment of Technical Skill (BOSATS), and GERT developed their tool for procedures, such as laparoscopic nissen fundoplication(LNF), where LS is an important skill that is applied. All of them used expert judgment. TAPP-CL and BOSATS also used task analysis and consensus methods. In simulation, the Objective Component Rating Scale (OCRS) and 'Laparoscopic Nissen Fundoplication-Checklist (LNF-CL)' were developed to assess simulated procedures such as LNF. OCRS was developed with expert judgment and consensus methods, and LNF-CL was developed using task analysis.

Response Process

For assessment of LS skills specifically, in the OR, the original Van Sickle LS error rating had structured rater training before starting the assessment process. In simulation, OSATS-GRS (original, Bingener 2008), GOALS (Stelzer 2009), original LS Checklist by Moorthy, and Van Sickle LS error rating (original, Takazawa 2015) had structured rater training before starting the assessment process.

For assessment of procedures that require LS, in the OR, original OSATS-GRS, LNF-CL and GERT had structured rater training before starting the assessment process. In simulation, none of the 9 tools had this type of validity evidence.

Internal Structure

For assessment of LS skills specifically, in the OR, all three; OSATS-GRS (Antosh 2013), original GOALS, and original Van Sickle LS error rating had evidence of rater reliability. In simulation, 20 of the 23 tools had evidence, the most common being rater reliability.

For assessment of procedures that require LS, in the OR, 15 of 17 tools provided evidence, the most common being rater reliability. In simulation, 6 of 9 showed evidence, with all using rater reliability.

Relations to Other Variables

For assessment of LS skills specifically, in the OR, none of the 3 tools reported this type of evidence. In simulation, 19 of 23 had evidence, the most common being the relationship to training level or case experience.

For assessment of procedures that require LS in the OR, 13 of 17 tools showed evidence, the most common being relationship to training level or case experience. Original GOALS has evidence using 3 methods. In simulation, 6 of 9 had evidence, all using relation to training level or case experience, and 'Assessment of Laparoscopic Roux-en Y gastric bypass (ALRYGB)' also using another method as well.

Consequences

For assessment of LS skills specifically, in the OR, none of the tools addressed this. In simulation, modified GOALS (Stelzer 2009), ASVC (King 2015), and LS

Checklist by Moorthy (Tijam 2013) gathered this type of evidence by developing a criterion-referenced score (benchmark or pass/fail).

For assessment of procedures that require LS in the OR or simulation, none of the tools addressed the issue of consequences.

Guidelines for usage

Through the analysis of all the included literature, Tables 5 and 6 (for tools used in the OR), and Appendix 4.2.5 (for tools used in the simulation setting) provide recommendations based on this review for selecting assessment tool that meets the needs of specific assessment conditions, along with a grading of the evidence within each condition (total score 1-5 limited evidence, 6-10 moderate evidence, and 11-15 strong evidence). The types of suturing and knot-tying techniques that were used for each assessment tool during the validation process is available in Appendix 4.2.2.

Discussion

This review allowed us to identify assessment tools that have been used in the OR/simulation to assess procedures that require LS or used in the OR/simulation to specifically assess LS skills, and to provide a summary of the validity evidence available for the tools through the contemporary framework of validity. We then provided guidelines for the selection of the appropriate tool depending on the assessment conditions. Overall, the assessment tools had the most evidence in internal structure and relations to other variables, moderate evidence in content, and response process, and the least amount of evidence in consequences. Nonetheless, for content validity, some tools could have been developed in studies that were not included in this review, such as the original GOALS and OSATS-GRS[125, 193]. Furthermore, when grading the evidence

for the assessment tools out of 15, none of them had a score of above 8 (moderate validity evidence), and most had limited validity evidence (a score of 1-5). Therefore, even though we have identified 42 assessment tools, most have limited validity evidence for their specified assessment condition, which is something that needs to be addressed in future studies.

The 'Criteria for Rating Validity Evidence' by Ghaderi et al that was used allowed us to summarize and score the validity evidence for an assessment tool that was provided from multiple studies. However, one of the aspects of the scoring that should be mentioned is the fact that it does not take into account the strength of the evidence for a particular source of validity. For instance, two tools were provided evidence in internal structure through rater reliability; however, one had a higher rater consistency than the other. This is something that surgical educators should pay attention to when using this rating criteria.

In terms of the assessment tools that were used to assess a procedure that require LS, some were generic to laparoscopy, some were specific to the procedure, and some had items that assessed suturing along with other items for that procedure (this information is provided in table 2). Therefore, the scores were a reflection of the overall performance of the trainee, rather than their LS performance specifically. None of the studies compared how scores correlate when assessing suturing portion of the procedure versus overall procedure. However, I think it is important to keep in mind that LS is not just a technical skill, but the performance of LS requires trainees to understand depth perception, dexterity, tissue handling etc, which are important skills in laparoscopy. Therefore, even though no study compared if assessing only LS portion correlates with

the overall performance, we cannot disregard the fact that there are skills needed during LS performance that applies to laparoscopy in general, which would be reflected in the overall performance. This is something that we cannot conclude from the included papers, and therefore, further studies would be required.

Researchers must pay attention to the fact that validity depends on the intended usage and the setting in which the assessment will be conducted in. This is why validity is a relative term, and validity evidence should be demonstrated in accordance with the assessment conditions; types of raters, direct versus recorded assessment, OR versus simulation among others.

When looking at the types of laparoscopic suturing and knot-tying that were used, we see that for assessment of LS skills, in simulation and in the OR, most used hand-sewn interrupted suturing, with intracorporeal knot-tying. However, for some tools, the type was not specified, hence, we cannot know which type of suturing was used, which limits the quality of the validity evidence. For assessment of procedures that require LS, in the OR and simulation, there was a mixture of different methods used. Some, however, did not specify the type of suturing and knot-tying used, or if it was handsewn or device assisted, which, again limits the generalizability of the findings.

By following this, if a training program will be using an assessment tool in the OR setting with the attending surgeon as the rater, to specifically assess LS skills, then using the guide that we have developed, they should select the tool or tools that have evidence for those specific conditions. For example, the original LS Checklist by Moorthy, which was most commonly used in simulation, has evidence to specifically assess LS skills for recorded performances in the simulation setting only. Therefore, if

this tool were to be used in the OR or for direct assessment, validity evidence should be collected before usage. In addition, this tool was used for assessing hand-sewn interrupted suturing, with intracorporeal knot-tying.

Another example is the original GOALS. This has evidence to specifically assess LS skills in both OR (recorded video) and simulation (recorded video, direct with observer). It also has evidence to assess procedures that require LS, only in the OR (direct with observer, attending, and self). Therefore, if GOALS were to be used to assess procedures that require LS in simulation, validity evidence should be collected before usage, taking into account the conditions of assessment, which include rater type and recorded video versus direct assessment. Hence, if an assessment tool will be implemented in a training program, or used for a specific purpose, validity evidence for that specific condition should be demonstrated. The type and amount of evidence needed will depend on the intended consequences of the assessment. However, providing validity evidence for all the different assessment conditions for all of the 42 assessment tools is not feasible. There should be a consensus among the leaders in education regarding the selection of one. This way, it would be more feasible to focus on providing validity evidence for that selected assessment tool in various assessment conditions, which would help in having not only limited, but strong validity evidence for the tool.

Conclusion

In conclusion, this review identifies assessment tools that were used in the OR/simulation to assess procedures that require LS in the OR/simulation or to specifically assess LS skills. Using the contemporary framework, the validity evidence available for these tools under various conditions is summarized. The results of this

review could be used as a guideline when implementing assessment tools into training programs.

Table 1: Study characteristics (N=68)

Characteristics	**N(%)**
Country	
US	28(41)
Canada	11(16)
UK	6(9)
Spain	4(7)
Japan	4(6)
Others*	15(21)
Year of publication	
2000-2008	20(29)
2009-2017	48(71)
Study design	
Development and/or validation of the tool	24(35)
For educational intervention	34(50)
For both	10(15)

*3 Chile, 4 Netherlands, 2 France, 2 Ireland, , 1 Germany, 1 Israel, 1 Turkey

Table 2: Characteristics of the assessment tools

	Type of items	Total number of items (LS specific items)	Total score	Setting								
				Recorded	Direct observation			Location		Simulation		OR
				Reviewer	Observer	Attending	Self	OR	Simulation	Type	Procedure	Procedure
<u>Global Rating Scale</u> OSATS-GRS Original [138, 151, 152, 154-158, 174, 178]	Generic	7(0)	35	+	+	+		+	+	Live porcine, box#	LS, LNF, LH	LTG, JJ-LRYGB
Dath 2004[147]	Generic	7(0)	35	+					+	Live porcine	LNF	
Hiemstra 2011[159]	Generic	6(0)	30			+		+				LH
Crochet 2016[146]	Generic	6(0)	30	+					+	Box	LH	
Birkmeyer 2013[137]	Generic	5(0)	25	+				+				LRYGB
Bingener 2008[136, 141, 171]	Generic	5(0)	25	+		+	+	+	+	Box	LS	HM, LRYGB
Broe 2006[73, 144, 150]	Generic	5(0)	25	+					+	Box	LS	
Kowalewski 2014[165]	Generic	2(0)	10	+					+	Box	LS	
Antosh	Hybri	10(4)	50	+				+				LS

2013[3]	d[c]											
GOALS Original [3, 32, 135, 154, 161, 164, 167, 184, 189]	Generic	5(0)	25	+		+	+	+	+	Box	LS	LS, HM, LRYGB, LNF, LIHR, UGI
Stelzer 2009[51, 163, 169, 179, 180]	Generic	4(0)	20	+					+	Live porcine, box, cadaver	LS	
Lee 2012[170]	Hybrid[c]	7(1)	35	+				+				LU
GOALS-GH[167, 177]	Hybrid[d]	5(0)	25	+	+	+	+	+				LIHR
LSIKT-GRS[145, 162]	Generic	4(0)	20	+	+				+	Box	LS	
LVG[191]	Generic[a]	1(0)	5	+					+	Box	LS	
GRS-OS[132, 142, 188]	Generic	5(0)	25	+				+	+	Live porcine, box	JJ-LRYGB	JJ-LRYGB
ALRYGB[132, 142, 143, 154, 188]	Specific[f]	4(2)	20	+		+	+	+	+	Live&cadaveric porcine, box	JJ-LRYGB	JJ-LRYGB
OCRS Original [147, 154]	Specific[f]	7(1)	35	+		+	+	+	+	Live porcine	LNF	LNF
Ghaderi 2015[154]	Specific[f]	10(1)	50			+	+	+				HM
ASVC[161]	Hybrid[c]	8(3)	40	+					+	Box	LS	
ASVC[184]	Hybrid[c]	7(3)	35	+					+	Box	LS	
LS-GRS[192]	Hybrid[c]	14(10)	70	+					+	Box	LS	
LS-	Hybri	3(1)	30	+					+	Box	LS	

QRS[192]	d[c]									
LS-AR[140]	Specific[e]	7(7)	35	+			+	Box	LS	
FLP[131]	Generic	4(0)	16	+			+	Box	LIHR	
BOSATS[31, 138]	Specific[f]	47(8)	195	+		+				LRYGB, JJ-LRYGB
GRITS[149]	Generic	9(0)	45		+	+				LIHR
OPRs[168]	Hybrid[d]	9(0)	45	+		+				LIHR
Checklist **LS Checklist by Moorthy** Original [133, 150, 166, 171, 172, 181, 183]	Specific[e]	29(29)	29	+	+		+	Box	LS	
Munz 2007[68, 148, 173]	Specific[e]	27(27)	27	+			+	Box	LS	
Tjiam 2013[182]	Specific[e]	3(3)	3	+			+	Box	LS	
IKT-CL[134]	Specific[f]	21(11)	1) 21		+		+	Live porcine	LS	
LS-CL[134]	Specific[f]	14(12)	2) 14		+		+	Live porcine	LS	
LSIKT-CL[145, 162]	Specific[e]	13(13)	13	+	+		+	Box	LS	
LNF-CL[175, 176]	Specific[f]	65(10)	65	+		+	+	Live porcine	LNF	LNF
TAPP-CL[177]	Hybrid[d]	24(2)	24	+		+				LIHR
ASVC[190]	Specific[e]	13(13)	13	+			+	Box	LS	
Error rating scale **Van Sickle LS error rating**										

Original [185-187]	Specific[e]	11(11)	N/A[b]	+		+	+	Live porcine	LS of LNF	LS of LNF
Takazawa 2015[181]	Specific[e]	9(9)	N/A	+			+	Box	LS	
LS-ERS[153]	Specific[e]	14(14)	N/A	+			+	Box	LS	
LIHR-ERS[131]	Specific[f]	6(2)	N/A	+			+	Box	LIHR	
GERT[138, 139, 160]	Generic	9(1)	N/A	+		+				LRYGB, LH, JJ-LRYGB

LS Laparoscopic Suturing, LTG Laparoscopic total gastrectomy, LH Laparoscopic hysterectomy, LRYGB Laparoscopic Roux-en Y gastric bypass, LU Laparoscopic urology cases requiring suturing and knot tying, LIHR Laparoscopic inguinal hernia repair, JJ-LRYGB Jenunojejunostomy-LRYGB, LNF Laparoscopic nissen fundoplication, HM Heller Myotomy, UGI Upper gastrointestinal.

OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, GOALS Global Operative Assessment of Laparoscopic Skills, GOALS-GH Groin Hernia, LSIKT LS and intracorporeal knot tying, LVG Laparoscopic Video Grader, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of LRYGB, OCRS Objective Component Rating Scale, ASVC Assessment for suturing of vaginal cuff, LS-QRS LS-Quality rating scale, LS-AR LS in augmented reality simulator, FLP Fundamentals to laparoscopic procedures, BOSATS Bariatric Objective Structured Assessment of Technical Skill, GRITS Global Rating Index for Technical Skills, OPRs Operative Performance Rating System, IKT Intracorporeal knot tying, LNF-CL Laparoscopic nissen fundoplication-Checklist, TAPP-CL Transabdominal Preperitoneal procedure-Checklist, LS-ERS Laparoscopic suturing-Error rating scale, LIHR-ERS Laparoscopic inguinal hernia repair-Error rating scale, GERT Generic Error Rating Tool.

Box simulation includes ex-vivo, virtual and augmented reality simulators

[a]No scale descriptions

[b]Error counting

[c]Hybrid: Generic and LS specific

[d]Hybrid: Generic and procedure specific

[e]Specific: Specific to LS

[f]Specific: Specific to a procedure

Table 3: Validity evidence for the assessment tools that were used to **specifically assess laparoscopic suturing in the operating-room**. Each evidence category score is out of 3, with a total score of 15.

		Content	Response Process	Internal Structure	Relations to Other Variables	Consequences	Total
OSATS-GRS	Antosh 2013	1	0	1	0	0	**2**
GOALS	Original	0	0	1	0	0	**1**
Van Sickle LS error rating	Original	2	2	1	1	0	**6**

OSATS-GRS Objective Structured Assessment of Technical Skills-Global rating scale, GOALS Global Operative Assessment of Laparoscopic Skills, LS Laparoscopic Suturing.

Table 4: Validity evidence for the assessment tools that were used to **assess procedures that required laparoscopic suturing in the operating-room** (e.g. laparoscopic nissen fundoplication). The procedures are stated in table 2. Each evidence category score is out of 3, with a total score of 15.

		Content	Response Process	Internal Structure	Relations to Other Variables	Consequences	Total
OSATS-GRS	Original	1	1	1	1	0	4
	Birkmeyer 2013	1	0	2	0	0	3
	Bingener 2008	1	0	1	0	0	2
	Hiemstra 2011	1	0	0	1	0	2
GOALS	Original	1	1	3	3	0	8
	Lee 2012	1	0	1	2	0	4
	GOALS-GH	2	0	2	3	0	7
	GRS-OS	1	0	0	1	0	2
	ALRYGB	1	0	1	1	0	3
	LNF-CL	2	1	1	0	0	4
	GRITS	2	0	1	1	0	4
	TAPP-CL	3	0	1	2	0	6
OCRS	Original	1	0	1	0	0	2
	Ghaderi 2015	1	0	1	0	0	2
	OPRs	1	0	1	0	0	2
	BOSATS	3	0	2	3	0	8
	GERT	3	1	1	3	0	8

OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, GOALS Global Operative Assessment of Laparoscopic Skills, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of Laparoscopic Roux-en Y gastric bypass, BOSATS Bariatric Objective Structured Assessment of Technical Skill, LNF-CL Laparoscopic nissen fundoplication-Checklist, GRITS Global Rating Index for Technical Skills, GERT Generic Error Rating Tool, TAPP-CL Transabdominal Preperitoneal procedure-Checklist, OCRS Objective Component Rating Scale, GOALS-GH Groin Hernia, OPRs Operative Performance Rating System

Table 5: OR setting: Guideline for the selection of an assessment tool that was used to **specifically assess laparoscopic suturing skills.**

			GRS	Checklist	Error rating scale
Recorded	**Reviewer**	**Generic**	**GOALS**(original)*		
		Specific			**Van Sickle LS error rating** (original)**
		Hybrid	**OSATS**(Antosh 2013)*		

GRS Global rating scale, LS Laparoscopic Suturing, OSATS Objective Structured Assessment of Technical Skills, GOALS Global Operative Assessment of Laparoscopic Skills
*limited (1-5) **moderate (6-10) ***strong level of validity evidence (11-15)

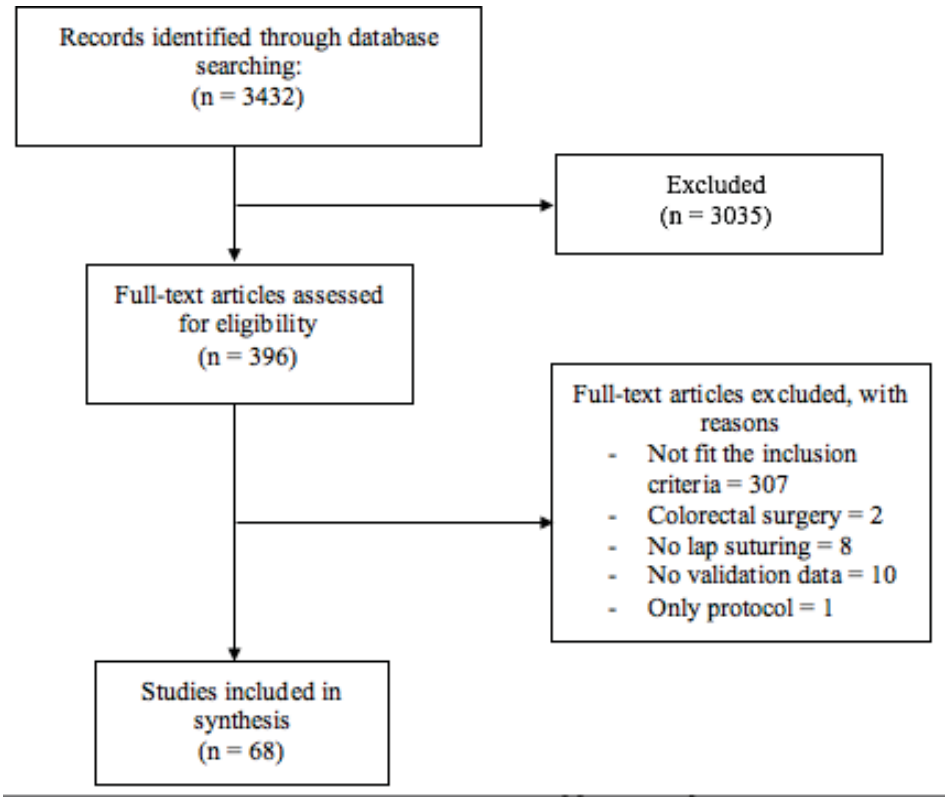
Table 6: OR setting: Guideline for the selection of an assessment tool that was used to **assess procedures that required laparoscopic suturing in the operating-room** (e.g. laparoscopic nissen fundoplication).

			GRS	Checklist	Error rating scale
Recorded	**Reviewer**	**Generic**	**OSATS**(original*, Birkmeyer 2013*)		
			GRS-OS*		
		Specific	**ALRYGB***		
			BOSATS**	**LNF-CL***	**GERT****
		Hybrid	**OPRs***		
			GOALS-GH*		
			GOALS(Lee 2012)*	**TAPP-CL****	
Direct	**Observer**	**Generic**	**GOALS**(original)*		
		Hybrid	**GOALS-GH****		
	Attending	**Generic**	**OSATS**(original*, Hiemstra 2011*, Bingener 2008*)		
			GOALS (original)**		
			GRITS*		
		Specific	**ALRYGB***		
			OCRS(original*, Ghaderi 2015*)		
		Hybrid	**GOALS-GH****		
	Self	**Generic**	**GOALS** (original)*		
			OSATS(Bingener 2008)*		
		Specific	**ALRYGB***		
			OCRS(original*, Ghaderi 2015*)		
		Hybrid	**GOALS-GH****		

GRS Global rating scale, LS Laparoscopic Suturing, OSATS Objective Structured Assessment of Technical Skills, GOALS Global Operative Assessment of Laparoscopic Skills, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of Laparoscopic Roux-en Y gastric bypass, BOSATS Bariatric Objective Structured Assessment of Technical Skill, LNF-CL Laparoscopic nissen fundoplication-Checklist, GRITS Global Rating Index for Technical Skills, GERT Generic Error Rating Tool, TAPP-CL Transabdominal Preperitoneal procedure-Checklist, OCRS Objective Component Rating Scale, GOALS-GH Groin Hernia, OPRs Operative Performance Rating System

*limited (1-5) **moderate (6-10) ***strong level of validity evidence (11-15)

Figure 1: Study identification and selection flow chart



4.3 Reliable Assessment of Operative Performance

Running Head: Operative Performance Assessment

Authors:
Elif Bilgic[1]BSc, Yusuke Watanabe[1,2] MD, Katherine M. McKendy[1] MD, Amani Munshi[1] MD, Yoichi M. Ito[3] PhD, Gerald M. Fried[1] MD, Liane S. Feldman[1] MD, Melina C. Vassiliou[1] MD

Institutions and Affiliations:
1. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada
2. Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan
3. Department of Biostatistics, Hokkaido University Graduate School of Medicine

Conflicts of Interest and Sources of Funding:
All of the authors do not have relevant conflicts of interest or financial ties to disclose. Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Covidien Canada. There was no funding or financial support for this work.

Corresponding author
Melina C. Vassiliou, MD
Associate Professor of Surgery
McGill University Health Centre
Montreal General Hospital
1650 Cedar Avenue, L9-313
Montreal, QC, H3G 1A4
melina.vassiliou@mcgill.ca

Abstract:

Background:
There is no consensus regarding the number of intraoperative assessments required to reliably measure trainee performance. This study used Generalizability Theory to describe factors contributing to score variance and to estimate the number of assessments needed to achieve high standards of reliability.

Methods:
While performing laparoscopic procedures, trainees were assessed by the attending surgeon using Global Operative Assessment of Laparoscopic Skills (GOALS). Data were collected prospectively (2-month intervals), assessing each trainee multiple times. Reliability coefficient was calculated using trainees, cases, and raters as factors.

Results:
Eighteen trainees were included for a total of 65 assessments. Total variance in scores was accounted for as follows: 66.1% by Trainees, 31.6% by the interaction between trainees and cases and 2.3% by raters. At least 3 cases are required for reliable scores using GOALS.

Conclusion:
Trainees accounted for most of the variance in GOALS scores with a minimum of 3 cases required to improve the reliability of the scores obtained. These data may guide the implementation of performance assessments in surgical training programs.

Key words: Generalizability theory; Reliability; Surgery; Assessment

Background

The way we train and assess surgeons has been evolving from case numbers and in-training evaluations to direct observations of performance using work-place based assessments. Various tools and instruments to measure operative performance are available to document that surgical trainees have achieved proficiency in a certain task or procedure[11]. The General Surgery Milestone Committee has recommended regular operative performance assessments as milestones for all general surgery residents[194]. However, their practical application in residency programs is still not well-established, and for most of them, little evidence is available on how to make decisions based on scores obtained using these tools. In order for these metrics to be used for trainee assessment, they must be reliable ie, the score should be consistent when the same trainee is assessed under the same conditions (assuming that the trainee's skills are stable). It is essentially impossible, however, to create the same conditions in the operating room from one case to another. Apart from trainee skill, scores may be affected by whether there is an easy or hard rater, the difficulty of the particular case, and the type of procedure. Furthermore, these factors may all be interacting simultaneously and can significantly impact scores, and put the reliability of a single evaluation into question, especially if the score is going to be used to make decisions about promotion or remediation of trainees.

Inter-rater reliability (raters), test retest reliability (cases), as well as other so-called "classic" methods are commonly used to assess reliability. Even though these are very useful, they have some limitations since the impact of raters or cases on scores is evaluated separately, and interactions between these factors cannot be taken into account. Generalizability theory (GT) is a statistical method in which the different factors contributing to variations in assessment scores and their interactions are taken into account when estimating reliability. GT permits the

integration of multiple factors that might simultaneously impact a trainee's score, into one reliability coefficient[195]. Furthermore, for a given assessment tool, once the overall reliability is calculated using GT, the number of raters or cases needed to reliably measure a trainee's skill level can be estimated[196].

The Global Operative Assessment of Laparoscopic Skills (GOALS) was developed to measure basic, generic laparoscopic skills and has been used to evaluate residents by direct observation in the operating room in multiple different studies and under various conditions[197, 198]. The initial publications on GOALS reported excellent inter-rater reliability for different raters assessing residents removing the gallbladder from the liver bed[125]. No study to date, however, has used GT to assess the reliability of GOALS scores obtained for trainees performing various procedures, and being assessed by various raters.

The purposes of this study were to apply GT: (1) to examine the impact of trainees, cases and raters on assessment scores using GOALS, (2) to determine the reliability coefficient of GOALS scores for one assessment by one rater, and (3) to evaluate the number of cases needed to obtain reliable GOALS scores.

Methods

Setting

This prospective study was conducted from July 2014 to January 2015 and was approved by the local Ethics Review Board of McGill University. General surgery residents at all levels and fellows were included. Using GOALS, trainees were assessed by the attending surgeon after each case. Trainees were assessed on multiple occasions within a 2-month interval.

Instrument

The Global Operative Assessment of Laparoscopic Skills (GOALS) is an assessment tool designed to measure basic laparoscopic skills, and is reported in detail elsewhere[125]. Briefly, it includes 5 domains: Depth Perception, Bimanual Dexterity, Efficiency, Tissue Handling, and Autonomy. Each domain is scored in a 5-point Likert scale with descriptive anchors at 1, 3 and 5. Scores range from 5-25. There is evidence supporting the validity of GOALS as a measure of generic laparoscopic skills when used for direct observation in the operating room. It has been used to measure skills in different institutions and over a wide range of both basic and advanced laparoscopic procedures[32, 197].

Rating Process

No additional training on how to use GOALS was provided to the attending surgeons, however all of them had experience using the tool in the past to assess resident performance in the operating room. The primary investigator was present in the operating room to provide the attending surgeons with a paper copy of the GOALS assessment tool at the end of every case. Attending surgeons were asked to complete the assessment immediately after the case. Assessments were only accepted if they were completed on the same day of the procedure.

Statistical Analysis

Generalizability Theory (GT) was used to determine the impact of the factors on assessment scores, and the overall reliability coefficient for the total GOALS score. Decision study (D Study) was then applied to determine the number of cases needed to reliably assess a trainee's skill level using GOALS[199]. Using JMP version 11 (SAS Institute Inc, Cary, NC), the variance of each component and the impact of each factor on assessment scores were calculated using Analysis of Variance (ANOVA), based on an unbalanced data set that was collected. The generalizability coefficient (overall reliability coefficient) and the number of cases

required were also calculated. Trainees (t), cases (c), and raters (r) were included as factors, along with their interaction terms (fully nested design)[200]. Cases and raters were labeled as random. The number of cases needed to achieve the recommended standards of a minimum reliability of 0.8 was determined[33].

Results

Eighteen trainees (3 PGY2, 1 PGY3, 3 PGY4, 8 PGY5 and 3 Fellows) underwent a median of 3 GOALS assessments (IQR 2-5) each (total of 65 assessments) by 9 attending surgeons. Ten raters participated in the study; however, one rater assigned almost perfect scores to all residents and was therefore excluded from the analysis. The laparoscopic procedures included cholecystectomies, adrenalectomies, colorectal cases, and hernia repairs (Table 1). Some of the procedures were infrequent, but they were not excluded since the frequency of the procedures cannot be controlled, and the case mix reflects the practice patterns in our setting.

The reliability coefficient for one assessment per trainee was 0.66. Trainee ability accounted for 66.1% of the variance in the assessment scores followed by the interaction between trainees and cases, which accounted for 31.6%, while raters accounted for 2.3%. Other factors and interactions had no effect. A summary of the %effect of factors is reported in Table 2. Increasing the number of cases per trainee assessed by a single attending increased the reliability of the GOALS assessment incrementally and a reliability of above 0.8 was achieved by 3 cases (Figure 1).

Discussion

Intraoperative tools to assess competency must have robust and transparent measurement properties. The present study suggests that for the GOALS tool, a single assessment by a single rater has a reliability coefficient of only 0.66, which is below the recommended standards[33].

Within a 2-month interval, more than 3 assessments per trainee were needed to provide a reliable assessment of laparoscopic skill (reliability coefficient of >0.8). This should be feasible in practice, since residents will very likely be involved in more than 3 or 4 laparoscopic procedures over that time frame.

However, it is important to note that the use of GOALS for formative assessment, geared at learning, might not require a reliability of above 0.8, which is a recommendation for summative assessments that might be used to make a decision about promotion to the next level, documentation of proficiency or attainment of a certain milestone. In addition, when used for summative assessment, the evaluations should be within a time frame where the learning of trainees is relatively stable so that the skill of the trainee can be reliably evaluated without the impact of their learning curve.

Classically, the reliability of scores of surgical performance has been estimated using methods such as inter-rater (comparing scores attributed by different raters), and test retest (assessing the consistency of scores between cases) reliability. These statistical methods can only estimate each of these potential sources of variance separately, and the simultaneous interactions between these factors, characteristic of the operating room environment, cannot be taken into account. Previous studies examined the inter-rater reliability of GOALS as a generic assessment tool using multiple different raters[125]. In addition, Fowler et al used GOALS to assess the performance of both novice and experienced trainees who performed both laparoscopic cholecystectomy and laparoscopic appendectomy (test-retest reliability), and no differences in GOALS scores were found between the 2 procedures, implying consistency in scores across procedures[201]. To date, no studies have used Generalizability theory to estimate the reliability of GOALS scores across different procedures. Generalizability theory incorporates the different

factors that might impact a trainee's score, and integrates them into one reliability coefficient and allows estimation of number of raters or cases needed for reliable assessment of trainee skill[196].

Other groups have identified the need for new methods to estimate reliability and have successfully applied GT to estimate the reliability of performance assessments in surgery. Williams et al used GT to determine the factors that affect the reliability of Operative Performance Rating scale (OPRs) scores and to determine the number of assessments needed to achieve a reliability above 0.8[202]. They included residents, raters and procedures as factors, and applied GT to each half-year of residency separately. They found that rater idiosyncrasies accounted for most of the variation in scores (36%) followed by resident ability (12%). In their study, procedures contributed to only 5% of the variance in OPRs scores. They concluded that 2.3 assessments per month per resident are required for reliable resident assessment using OPRs, resulting in a median of 18.5 assessments per year. Given the high impact of raters on scores, they recommended residents be rated by at least 10 raters per year, based on an earlier study done by their group[203].

In contrast, using a different assessment tool, we found that trainee skill accounted for the majority (66.1%) of the variance in scores, which is clearly desirable in an assessment tool of trainee skill. There was, however, a significant impact of trainee and case interaction on scores (31.6%), which is very important since this means that a resident operating on one case will have a different score than the same resident operating on another case regardless of the raters. There are many factors that might affect the level of difficulty of a case including patient variables, and the environment, so the resident's reaction to those factors cannot be disregarded. GOALS is a tool measuring fundamental laparoscopic skills that are transferable between cases. Therefore, a

resident may perform very well in an "easy" case and less well in a more challenging case; however, their overall generic laparoscopic skills should be consistent across different cases, if these factors can be accounted for. For this reason, residents should be assessed on a variety of different cases. Interestingly, and contrary to the study by Williams et al, there was very little variance in scores related to raters (2.3%), and in this study, none related specifically to case alone or to the interaction between rater and trainee. We suspect that the main reason for this difference in rater variance between our study and the study by Williams et al is that we applied GT to all levels of training, whereas they applied GT separately to each residency half-year. This way, our results had a lot more variance due to trainees since we included a range of levels from PGY1 to 5 and fellows. Williams et al assessed a group of residents in the same year, with similar skills and hence, their rater impact increased and trainee impact decreased. Our study also had a small size, and all of the surgeons were at the same institution and were already familiar with the GOALS instrument. The completion of the assessments immediately after the procedure may have also decreased the rater effect. Since the attending surgeons may have performed several surgeries in a given day or within a couple days, the assessments that are provided by the surgeons at a later time might not be reliable and since the study by Williams et al did not specify when the assessments were completed, this could have a significant impact. Also, the reliability coefficient might be different since OPRs and GOALS are not the same assessment tool. Reliability is specific to the instrument and testing conditions, and one coefficient cannot be compared to the other.

In our study, one rater assigned very high scores to all residents, so we excluded this rater's assessments from our analysis. When we analyzed the data including the assessments from this surgeon, the overall reliability decreased to 0.60 from 0.66. The impact of the raters

increased to 14.4% from 2.3% and the impact of cases decreased to 25.5% from 31.6%. The excluded rater was familiar with the assessment tool, but did not seem to use the range of the assessment tool when performing the assessments.

This preliminary study includes a small sample size and variety of procedures, somewhat limiting the generalizability of the results. In order to determine the reliability of GOALS as an assessment of fundamental skills across different procedures, a multi-center trial would have to be undertaken, including different geographic locations, and a variety of procedures and raters. We would then have a more accurate appreciation for the impact of various factors and interactions on scores and would be better equipped to make decisions based on the scores. Ideally, and depending on the consequences of performance assessments, this type of a study would need to be performed for each tool measuring different aspects of performance to provide guidance in a given training condition.

Conclusion

These preliminary data suggest that apart from trainee ability, the interaction between trainees and cases accounted for most of the variance in assessment scores. More than 3 assessments per trainee by the attending surgeon of the case within 2-month intervals, and ideally on different procedures, should be required to obtain reliable performance scores using GOALS. This methodology may be used to determine the number of assessments needed to provide reliable assessments of technical skills in laparoscopic surgery.

Table 1: Procedures included in the study

Procedures*	**N (%)**
Cholecystectomy	16 (25)
Colorectal Surgery	15 (23)
Ventral Hernia	9 (14)
UGI†	9 (14)
LIHR‡	6 (9)
Miscellaneous§	10 (15)

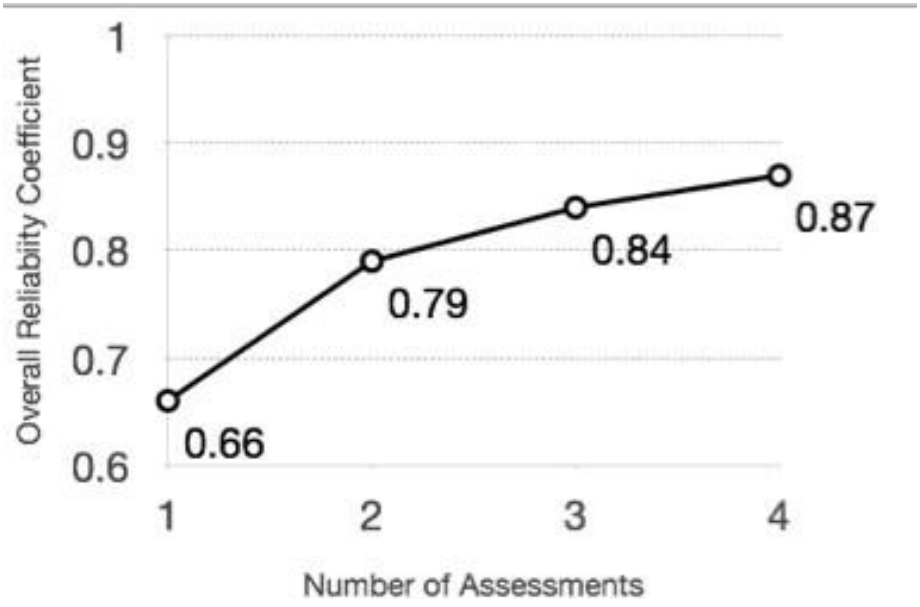
* Laparoscopic (Lap) procedures
† Upper gastrointestinal; Fundoplication, gastrectomy sleeve, herniorraphy paraesophegeal and heller myotomy with dor fundoplication
‡ Lap inguinal hernia repair
§ Lap adrenalectomy (6), Diagnostic laparoscopy (1), procedure unknown (3)

Table 2: Percent effect of each factor and their interactions on the performance score of the surgical trainees using GOALS.

Factor	% Effect
Trainees (t)	66.1
Cases (c)	0
Raters (r)	2.3
tc*	31.6
tr†	0
cr‡	0
tcr§	0

*tc: Trainee's tendency to score differently in one case versus the next regardless of the raters.
†tr: Rater's tendency to score a trainee differently regardless of the cases.
‡cr: Differences in the scores given by raters in one case versus the next regardless of the trainees.
§tcr: A rater's tendency to assess a trainee differently in one case versus the next.

Figure 1: Reliability of number of assessments per trainee assessed by a single attending surgeon using D study



CHAPTER 5: SUMMARY AND CONCLUSIONS

5.1 General Findings

Due to increased concerns over patient safety, training of surgical trainees is continuously evolving. The current paradigm focuses on learning outside of the clinical setting through simulation as an adjunct to learning in the operating room (OR). This training paradigm is called 'Competency-based Medical Education (CBME),' and is used for the training of laparoscopic skills, such as laparoscopic suturing (LS). Simulation training is an essential component of skill gain since trainees can use simulation to learn LS, and then start performing it in the OR with an already established skill-set on which to build and further refine their skills. This way, the initial part of the learning curve is shifted from the OR to the simulation setting where trainees can repeatedly practice LS, and there are no concerns over patient safety, time constraints, or costs associated with prolonged time spent in the OR for teaching. At the same time, even though simulation training is crucial, everything leads up to the performance of the trainees in the OR, and making sure that trainees are improving their LS skills to become competent surgeons. Within this training paradigm, assessment in both simulation and OR settings is important since we have to find ways to track trainee progress and make decisions about their readiness to move on to the next stage of their training. This work provides strategies for accurate assessment of fundamental and advanced LS skills in the simulation setting and the operating room.

As a first step, we conducted a literature review to identify simulation platforms that have been developed to assess LS skills (chapter 2). This review highlighted that most simulation platforms either target basic LS skills, such as the performance of a single suture, or they are not cost-effective (using live animal or ex-vivo models, which are expensive and hard to manage in terms of feasibility). However, various needs assessments identified that apart from a need for

strong training and assessment for basic LS skills, there was also a need for simulation platforms that target more advanced LS skills. Therefore in chapter 3, in order to create a comprehensive assessment platform including basic and advanced LS skills, we evaluated an existing simulation program called Fundamentals of Laparoscopic Surgery (FLS) that was identified in chapter 2 to accurately assess fundamental laparoscopic skills, including LS. Subsequently, we addressed the need for tasks that better reflect the complexities of the OR setting. This was done by developing and/or providing validity evidence for advanced LS tasks as measures of LS skills, for both free-needle and device-assisted suturing. The tasks were made from cost-effective and readily available materials, and target skills that were identified by Nepomnayshy et al. and Enani et al. to be important, such as suturing under tension, bowel anastamosis, device-assisted suturing, needle handling, and suturing in difficult angles[22, 23]. For the free-needle tasks, we added on the previous validity evidence by determining expert proficiency benchmarks that could be used by trainees as a goal when performing the LS tasks. For the automated-suturing task, we developed a task through expert opinion, and provided validity evidence in the context of skill assessment by showing that the task metrics could differentiate between different levels of surgeons, the scores are consistent between 2 raters, and has good correlation with Global Operative Assessment of Laparoscopic Skills (GOALS) assessment tool. Apart from that, expert participants found that the task is relevant for practice and could be used as a training mechanism.

Even though trainees could use these platforms to be competent in LS, we needed to make sure that the skill gain is transferable to the OR. Therefore, assessment of LS performance in the OR is the key to track trainee progress and to make sure that they are competent in performing LS in the clinical setting. Therefore, in the first part of chapter 4, we conducted an in

depth review of the literature in order to determine assessment tools that have either been developed and/or used in the validation process to assess LS skills specifically, or procedures that require LS in the OR. This way, we were able to identify in what assessment conditions the tools were used (who was the rater, was the assessment done directly or used video-recording, etc.), along with the validity evidence surrounding the assessment tools. In the end, we were able to develop a guideline for surgical educators in order to help them select an assessment tool that assesses LS. Most assessment tools had limited evidence, but one tool that had moderate evidence for assessment of advanced laparoscopic procedures that require LS through direct observation by the attending surgeon was GOALS. In the second part of chapter 4, we used the GOALS assessment tool and determined how much various factors that are not related to the trainee performance, such as raters and cases (including case difficulty and procedure type), affect the reliability (consistency) of the assessment scores of the trainees, and how many assessments we would need per trainee to minimize the effects of the external factors, so that we can accurately measure the trainee performance and trust that the score that they receive reflects their performance. We determined that the trainee performance itself had an effect of 66% on the assessment score, and the way trainees perform from case to case also had a significant effect on the score with 32%. This shows that in the OR, not only is the assessment score related to the trainee performance, but also it relates to case difficulty and procedure type. In order to minimize the effect of the cases on the score, we found that assessment in at least 3 cases are required per trainee to have reliable assessment scores. Chapter 4 addresses important points since we need to ensure that among all the assessment tools, surgical educators select the one that has the relevant validity evidence for their purposes and assessment conditions. Apart from that, educators need to know how to implement the tool into their programs with regards to how many assessments

provide an accurate representation of the trainee performance, so that we can make certain decisions about trainee progress and readiness.

5.2 Limitations

In this section, we will address some methodological limitations that should be considered when interpreting the results and conclusions of this thesis. In chapter 3, the first part included an analysis of the Fundamentals of Laparoscopic Surgery (FLS) exam results. In terms of the data, we only had data for 60% of the exam takers, which is a limitation since we do not know if the missing data was from a certain demographic or training level, which could have added some bias to our results. However, we have done a brief review to determine where the missing data might be, and there does not seem to be a pattern or bias emerging as a result of the missing data.

In the second part of chapter 3, we developed and provided validity evidence for advanced laparoscopic suturing (LS) tasks. For the free-needle tasks and the device-assisted suturing task, we did not use them as an educational intervention, determined if the trainees improved their LS skills after practicing on these tasks, or provided evidence for the transferability of the skill gain to the OR setting. Comparison studies with 2 groups (1 group trains until they achieve the pre-determined proficiency benchmarks, and the other group is a control without any interventions) could be done to determine if training with these tasks can improve skill in the simulation setting and in the OR. For the device-assisted suturing task specifically, we used one type of the automated-suturing device and asked participants to perform continuous suturing with intra-corporeal knot-tying. However, there are different types of automated suturing devices that are available, and surgeons can perform extra-corporeal knot-

tying with those devices as well. Additionally, the study was done at McGill University affiliated hospitals. Therefore, since we only have participants from McGill, our results might not generalize to other training programs where device-assisted suturing is used. Therefore, a multi-center approach, similar to the way it was done for the free-needle tasks, would allow participation of trainees from different programs, which would make our results more generalizable to other programs.

In chapter 4, in the first part, in order to summarize the validity evidence surrounding the assessment tools that were identified in our in-depth literature review, we used a rating criteria by Ghaderi et al. that was modified from Beckman et al[34, 204]. This rating criteria rates each of the 5 sources of validity from 0-3, with a total score of 15, with anchors explaining what assessors should look for and what each score means. Beckman et al. originally developed this tool based on the authors' expert opinions as to how to define a score of 0 versus a score of 3. Apart from that, this rating scale was not developed through a well-defined process, and the modifications that were done by Ghaderi et al. were again based on the authors' opinions without providing further justifications or evidence for the tool as a measure of validity. Even though this validity rating criteria has some limitations, there are no other criteria by which we could score and summarize the validity evidence surrounding the tools based on multiple studies, and this criteria provides a feasible way to do that.

5.3 Future Directions

The topics that were investigated in this thesis address some of the gaps in the literature, but also raise some valuable questions that could guide future research. The main focus of this thesis was laparoscopic suturing (LS), and usage of simulation and operating room (OR)

platforms for assessment purposes. Simulation tasks allow trainees to practice their LS skills repeatedly, getting to a certain point in their learning curve before performing in the OR. Therefore, using our advanced LS tasks as an educational tool is valuable and further research will shed light to how we can incorporate these platforms in a curriculum and show that practice with these tasks improves LS skill and that this improvement transfers to the OR setting.

Apart from that, our advanced LS tasks were all developed based on various needs assessments that identified LS as a skill that lacked simulation platforms to teach and assess advanced skills. We were able to identify this need in chapter 2 as well, where we found through our review that most simulations were for basic LS, such as a single suture, and/or they were using resource intensive and expensive animal models (in vivo and ex vivo). However, in the needs assessments, they identified other skills, such as dissection, retraction, and difficult exposure techniques, that should be also taught and assessed as a part of an advanced laparoscopic surgery curriculum. A program such as Fundamentals of Laparoscopic Surgery (FLS) not only focuses on basic LS skills but also focuses on fundamental skills required to perform laparoscopy (tissue handling, bimanual dexterity, etc.). It has been shown many times that training with the tasks of FLS improve skill and is transferable to the OR[19, 92, 121, 205]. Therefore, further research should focus on incorporation of our advanced LS tasks into an educational program with other tasks related to advanced skills, such as dissection, so that trainees can have a platform similar to FLS where they can train for advanced skills. Also, regarding our device-assisted suturing task, the same task could be used to teach and assess other forms of suturing and knot-tying techniques, along with using different types of devices that are available in the clinical setting.

Finally, in chapter 4, we identified many assessment tools that were either developed and/or used to assess LS specifically or a procedure that requires LS, and validity evidence for most of them was limited. I believe that there needs to be a consensus among surgical educators and experts in the field regarding which assessment tool should be used to assess LS specifically, and which assessement tool for a procedure that requires LS. This way, we can focus on the selected assessment tools, and provide high quality validity evidence for various assessment conditions. Otherwise, more studies will be done to develop assessment tools for LS with limited validity evidence.

5.4 Conclusions

In conclusion, I have investigated ways in which we can assess the laparoscopic suturing (LS) skills of the trainees in simulation and clinical settings. This thesis provides a starting point for future work aiming to improve assessment of trainees. We demonstrated how to use bench-top simulators and assessment tools in the operating room effectively to assess LS skills. This is crucial since technical skills are an important component of patient safety, and our efforts are aimed at making sure that trainees are competent in performing a technical skill, such as LS.

CHAPTER 6: LIST OF REFERENCES

1. Zendejas, B., R. Hernandez-Irizarry, and D.R. Farley, *Does Simulation Training Improve Outcomes in Laparoscopic Procedures?* Advances in Surgery, 2012. **46**(1): p. 61-71.
2. Seymour, N.E., et al., *Virtual reality training improves operating room performance: results of a randomized, double-blinded study.* Annals of surgery, 2002. **236**(4): p. 458-4.
3. Antosh, D.D., et al., *Blinded Assessment of Operative Performance After Fundamentals of Laparoscopic Surgery in Gynecology Training.* Journal of Minimally Invasive Gynecology, 2013. **20**: p. 353-359.
4. Birkmeyer, J.D., et al., *Surgical skill and complication rates after bariatric surgery.* New England Journal of Medicine, 2013. **369**(15): p. 1434-1442.
5. Beard, J.D., et al., *Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology.* Health Technol Assess, 2011. **15**(1): p. i-xxi, 1-162.
6. Hamstra, S.J. and A. Dubrowski, *Effective Training and Assessment of Surgical Skills, and the Correlates of Performance.* Surgical Innovation, 2005. **12**(1): p. 71-77.
7. Frank, J.R., et al., *Competency-based medical education: theory-to practice.* Medical Teacher, 2010. **32**: p. 638-645.
8. ten Cate, O. and F. Scheele, *Competency-based postgraduate training: can we bridge the gap between theory and clinical practice?* Academic medicine : journal of the Association of American Medical Colleges, 2007. **82**(6): p. 542-7.
9. Touchie, C. and O. Ten Cate, *The promise, perils, problems and progress of competency-based medical education.* Medical Education, 2016. **50**(1): p. 93-100.
10. Feldman, L.S., et al., *Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents.* Journal of American College of Surgeons, 2004. **198**: p. 105-110.
11. Vassiliou, M.C. and L.S. Feldman, *Objective assessment, selection, and certification in surgery.* Surgical Oncology, 2011. **20**(3): p. 140-145.
12. Jr, W.E.K., *The Evolution of Laparoscopy and the Revolution in Surgery in the Decade of the 1990s.* Journal of the Society of Laparoendoscopic Surgeons, 2008. **12**: p. 351-357.
13. Derossis, A.M., et al., *Development of a model for training and evaluation of laparoscopic skills.* American journal of surgery, 1998. **175**: p. 482-7.
14. Stefanidis, D., et al., *Simulation in surgery: what's needed next?* Annals of surgery, 2015. **261**(5): p. 846-53.
15. Scott, D.J., et al., *Laparoscopic training on bench models: better and more cost effective than operating room experience?* Journal of the American College of Surgeons, 2000. **191**(3): p. 272-83.
16. Fried, G.M., *Simulators for laparoscopic surgery- a coming of age.* Asian Journal of Surgery, 2004. **27**: p. 1-3.
17. Peters, J.H., et al., *Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery.* Surgery, 2004. **135**(1): p. 21-27.
18. Sroka, G., et al., *Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial.* American Journal of Surgery, 2010. **199**(1): p. 115-120.

19. McCluney, A.L., et al., *FLS simulator performance predicts intraoperative laparoscopic skill.* Surgical endoscopy, 2007. **21**(11): p. 1991-5.
20. Korndorffer, J.R., Jr., et al., *Simulator training for laparoscopic suturing using performance goals translates to the operating room.* Journal of the American College of Surgeons, 2005. **201**(1): p. 23-9.
21. Mattar, S.G., et al., *General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors.* Annals of surgery, 2013. **258**(3): p. 440-9.
22. Enani, G., et al., *What are the Training Gaps for Acquiring Laparoscopic Suturing Skills?* J Surg Educ, 2017.
23. Nepomnayshy, D., et al., *Identifying the need for and content of an advanced laparoscopic skills curriculum: Results of a national survey.* American Journal of Surgery, 2016. **211**(2): p. 421-425.
24. Lim, S., et al., *Laparoscopic Suturing as a Barrier to Broader Adoption of Laparoscopic Surgery.* JSLS, 2017. **21**(3).
25. Sadideen, H. and R. Kneebone, *Practical skills teaching in contemporary surgical education: how can educational theory be applied to promote effective learning?* Am J Surg, 2012. **204**(3): p. 396-401.
26. Stefanidis, D., et al., *Redefining simulator proficiency using automaticity theory.* Am J Surg, 2007. **193**(4): p. 502-6.
27. Medina, M., *Formidable Challenges to Teaching Advanced Laparoscopic Skills.* Journal of the Society of Laparoendoscopic Surgeons, 2001. **5**(2): p. 153-158.
28. Gallagher, A.G. and G.C. O'Sullivan, *Fundamentals of surgical simulation*, ed. P. Apell. 2012, London: Springer.
29. S Swaroop Vedula, Masaru Ishii, and G.D. Hager, *Objective computer-aided technical skill evaluation.* Annual review of biomedical engineering, 2017. **19**: p. 301-325.
30. Khan, R.S., et al., *Analysis of eye gaze: do novice surgeons look at the same location as expert surgeons during a laparoscopic operation?* Surg Endosc, 2012. **26**(12): p. 3536-40.
31. Zevin, B., et al., *Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery.* Journal of the American College of Surgeons, 2013. **216**: p. 931-955,1033.
32. Hogle, N.J., et al., *Evaluation of surgical fellows' laparoscopic performance using Global Operative Assessment of Laparoscopic Skills (GOALS).* Surgical Endoscopy and Other Interventional Techniques, 2014. **28**: p. 1284-1290.
33. Williams, R.G., et al., *Assuring the reliability of resident performance appraisals: More items or more observations?* Surgery, 2005. **137**(2): p. 141-147.
34. Ghaderi, I., et al., *Technical Skills Assessment Toolbox.* Annals of Surgery, 2015. **261**(2): p. 251-262.
35. Williams, R.G., M.J. Kim, and G.L. Dunnington, *Practice Guidelines for Operative Performance Assessments.* Ann Surg, 2016. **264**(6): p. 934-948.
36. Crossley, J., et al., *Generalisability: A key to unlock professional assessment.* Medical Education, 2002. **36**(10): p. 972-978.
37. Palter, V.N., et al., *Resident perceptions of advanced laparoscopic skills training.* Surgical endoscopy, 2010. **24**(11): p. 2830-4.

38. Rattner, D.W., K.N. Apelgren, and W.S. Eubanks, *The need for training opportunities in advanced laparoscopic surgery.* Surg Endosc, 2001. **15**(10): p. 1066-70.
39. Ghaderi, I., et al., *Technical skills assessment toolbox: a review using the unitary framework of validity.* Ann Surg, 2015. **261**(2): p. 251-62.
40. Downing, S., *Validity: on meaningful interpretation of assessment data.* Med Educ 2003. **37**: p. 830–837
41. Kroeze, S.G., et al., *Assessment of laparoscopic suturing skills of urology residents: a pan-European study.* Eur Urol, 2009. **56**(5): p. 865-72.
42. Kowalewski, T.M., et al., *Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology.* J Surg Res, 2014. **192**(2): p. 329-38.
43. Moorthy, K., et al., *Bimodal assessment of laparoscopic suturing skills: construct and concurrent validity.* Surg Endosc, 2004. **18**(11): p. 1608-12.
44. Zheng, B., et al., *Building an efficient surgical team using a bench model simulation: construct validity of the Legacy Inanimate System for Endoscopic Team Training (LISETT).* Surg Endosc, 2008. **22**(4): p. 930-7.
45. Boza, C., et al., *A cadaveric porcine model for assessment in laparoscopic bariatric surgery--a validation study.* Obes Surg, 2013. **23**(5): p. 589-93.
46. Lin, D.W., et al., *Computer-based laparoscopic and robotic surgical simulators: performance characteristics and perceptions of new users.* Surg Endosc, 2009. **23**(1): p. 209-14.
47. Broe, D., et al., *Construct validation of a novel hybrid surgical simulator.* Surgical Endoscopy and Other Interventional Techniques, 2006. **20**(6): p. 900-904.
48. Van Sickle, K.R., et al., *Construct validation of the ProMIS simulator using a novel laparoscopic suturing task.* Surgical Endoscopy and Other Interventional Techniques, 2005. **19**(9): p. 1227-1231.
49. Duffy, A.J., et al., *Construct validity for the LAPSIM laparoscopic surgical simulator.* Surg Endosc, 2005. **19**(3): p. 401-5.
50. Hennessey, I.A. and P. Hewett, *Construct, concurrent, and content validity of the eoSim laparoscopic simulator.* J Laparoendosc Adv Surg Tech A, 2013. **23**(10): p. 855-60.
51. Deal, S.B., et al., *Crowd-sourced assessment of technical skills: An opportunity for improvement in the assessment of laparoscopic surgical skills.* American Journal of Surgery, 2016. **211**: p. 398-404.
52. Pagador, J.B., et al., *Decomposition and analysis of laparoscopic suturing task using tool-motion analysis (TMA): improving the objective assessment.* Int J Comput Assist Radiol Surg, 2012. **7**(2): p. 305-13.
53. Palter, V.N., et al., *Designing a proficiency-based, content validated virtual reality curriculum for laparoscopic colorectal surgery: a Delphi approach.* Surgery, 2012. **151**(3): p. 391-7.
54. Chang, O.H., et al., *Developing an Objective Structured Assessment of Technical Skills for Laparoscopic Suturing and Intracorporeal Knot Tying.* J Surg Educ, 2016. **73**(2): p. 258-63.
55. Kowalewski, K.F., et al., *Development and validation of a sensor- and expert model-based training system for laparoscopic surgery: the iSurgeon.* Surgical Endoscopy and Other Interventional Techniques, 2016: p. 1-11.

56. Strickland, A., et al., *Development of an ex vivo simulated training model for laparoscopic liver resection.* Surgical Endoscopy and Other Interventional Techniques, 2011. **25**(5): p. 1677-1682.
57. Trejos, A.L., et al., *Development of force-based metrics for skills assessment in minimally invasive surgery.* Surgical Endoscopy, 2014. **28**: p. 2106-2119.
58. Stefanidis, D., et al., *Establishing technical performance norms for general surgery residents.* Surg Endosc, 2014. **28**(11): p. 3179-85.
59. Lusch, A., et al., *Evaluation of the impact of three-dimensional vision on laparoscopic performance.* J Endourol, 2014. **28**(2): p. 261-6.
60. Escamirosa, F.P., et al., *Face, content, and construct validity of the EndoViS training system for objective assessment of psychomotor skills of laparoscopic surgeons.* Surg Endosc, 2015. **29**(11): p. 3392-403.
61. Steinemann, D.C., et al., *Internal retraction in single-port laparoscopic cholecystectomy: initial experience and learning curve.* Minim Invasive Ther Allied Technol, 2013. **22**(3): p. 171-6.
62. Romero, P., et al., *Intracorporal suturing--driving license necessary?* J Pediatr Surg, 2014. **49**(7): p. 1138-41.
63. Hiemstra, E., et al., *Intracorporeal suturing- economy of instrument movements using a box trainer model.* Journal of Minimally Invasive Gynecology, 2011. **18**(4): p. 494-499.
64. Leeds, S.G., et al., *Learning Curve Associated With an Automated Laparoscopic Suturing Device Compared With Laparoscopic Suturing.* Surgical Innovation, 2017. **24**(2): p. 109-114.
65. Botden, S.M.B.I., I.H.J.T.d. Hingh, and J.J. Jakimowicz, *Meaningful assessment method for laparoscopic suturing training in augmented reality.* Surgical Endoscopy, 2009. **23**: p. 2221-2228.
66. Zheng, B., et al., *Measuring mental workload during the performance of advanced laparoscopic tasks.* Surg Endosc, 2010. **24**(1): p. 45-50.
67. Stefanidis, D., et al., *Multicenter longitudinal assessment of resident technical skills.* Am J Surg, 2015. **209**(1): p. 120-5.
68. Sánchez-Margallo, J.A., et al., *Objective assessment based on motion-related metrics and technical performance in laparoscopic suturing.* International Journal of Computer Assisted Radiology and Surgery, 2016: p. 1-8.
69. Yamaguchi, S., et al., *Objective assessment of laparoscopic suturing skills using a motion-tracking system.* Surg Endosc, 2011. **25**(3): p. 771-5.
70. Dubrowski, A., et al., *Quantification of process measures in laparoscopic suturing.* Surgical Endoscopy, 2006. **20**: p. 1862-1866.
71. Dayan, A.B., et al., *A Simple, Low-cost Platform for Basic Laparoscopic Skills Training.* Surgical Innovation, 2008. **15**(2): p. 136-142.
72. Figert, P.I., et al., *Transfer of training in acquiring laparoscopic skills.* J Am Coll Surg., 2001. **193**(5): p. 533-537.
73. Buckley, C.E., et al., *Zone calculation as a tool for assessing performance outcome in laparoscopic suturing.* Surgical Endoscopy and Other Interventional Techniques, 2015. **29**: p. 1553-1559.
74. Poursartip, B., et al., *Analysis of Energy-based Metrics for Laparoscopic Skills Assessment.* IEEE Trans Biomed Eng, 2017.

75. Botden, S.M., et al., *Augmented versus virtual reality laparoscopic simulation: what is the difference? A comparison of the ProMIS augmented reality laparoscopic simulator versus LapSim virtual reality laparoscopic simulator.* World J Surg, 2007. **31**(4): p. 764-72.
76. Kobayashi, S.A., et al., *Bringing the skills laboratory home: an affordable webcam-based personal trainer for developing laparoscopic skills.* J Surg Educ, 2011. **68**(2): p. 105-9.
77. O'Neill Trudeau, M., et al., *Construct validity and educational role for motion analysis in a laparoscopic trainer.* Surgical Endoscopy, 2015. **29**: p. 2491-2495.
78. Sharma, M., et al., *Construct validity of fresh frozen human cadaver as a training model in minimal access surgery.* JSLS, 2012. **16**(3): p. 345-52.
79. Kowalewski, K.-F., et al., *Development of a sensor-and expert model based training device for suturing and knot tying in laparoscopic surgery.* Surgical Endoscopy, 2017. **31**: p. 2155-2165.
80. Horeman, T., et al., *Force measurement platform for training and assessment of laparoscopic skills.* Surgical Endoscopy and Other Interventional Techniques, 2010. **24**(12): p. 3102-3108.
81. Rosser, J.C., Jr., et al., *Impact of Super Monkey Ball and Underground video games on basic and advanced laparoscopic skill training.* Surg Endosc, 2017. **31**(4): p. 1544-1549.
82. Zdichavsky, M., et al., *Laparoscopic gastro-jejunal anastomosis using novel r2 deflectable instruments in an ex vivo model.* Minim Invasive Ther Allied Technol, 2016. **25**(2): p. 91-8.
83. Uemura M, Y.M., Tomikawa M, Obata S, Souzaki R, Ieiri S, Ohuchida K, Matsuoka N, Katayama T, Hashizume M., *Objective assessment of the suture ligature method for the laparoscopic intestinal anastomosis model using a new computerized system.* Surg Endosc., 2015. **29(2)**(1432-2218 (Electronic)): p. 444-52.
84. Veneziano, D., et al., *Preliminary evaluation of the SimPORTAL major vessel injury (MVI) repair model.* Surg Endosc, 2016. **30**(4): p. 1405-12.
85. Keyser, E.J., et al., *A simplified simulator for the training and evaluation of laparoscopic skills.* Surgical Endoscopy, 2000. **14**(2): p. 149-153.
86. Xeroulis, G., A. Dubrowski, and K. Leslie, *Simulation in laparoscopic surgery: a concurrent validity study for FLS.* Surg Endosc, 2009. **23**(1): p. 161-5.
87. Bahsoun, A.N., et al., *Tablet based simulation provides a new solution to accessing laparoscopic skills training.* J Surg Educ, 2013. **70**(1): p. 161-3.
88. Oostema, J.A., M.P. Abdel, and J.C. Gould, *Time-efficient laparoscopic skills assessment using an augmented-reality simulator.* Surgical Endoscopy, 2008. **22**: p. 2621-2624.
89. Sleiman, Z., et al., *Validation Study of a Portable Home Trainer Using a Pad for Laparoscopic Practice.* Surgical Innovation, 2017. **24**(3): p. 284-288.
90. Yeung, C., et al., *Video assessment of laparoscopic skills by novices and experts: implications for surgical education.* Surg Endosc, 2017. **31**(10): p. 3883-3889.
91. Vassiliou, M.C., et al., *The MISTELS program to measure technical skill in laparoscopic surgery : evidence for reliability.* Surg Endosc, 2006. **20**(5): p. 744-7.
92. Ritter, M.E. and D.J. Scott, *Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic Surgery.* Surgical Innovation, 2007. **14**(2): p. 107-112.
93. Fried, G.M., *FLS assessment of competency using simulated laparoscopic tasks.* Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract, 2008. **12**(2): p. 210-2.

94. Fried, G.M., et al., *Proving the Value of Simulation in Laparoscopic Surgery.* Annals of Surgery, 2004. **240**(3): p. 518-528.
95. Mason, J.D., et al., *Is motion analysis a valid tool for assessing laparoscopic skill?* Surgical Endoscopy, 2013. **27**(5): p. 1468-1477.
96. Chmarra, M.K., C.A. Grimbergen, and J. Dankelman, *Systems for tracking minimally invasive surgical instruments.* Minim Invasive Ther Allied Technol, 2007. **16**(6): p. 328-340.
97. Pugh, C.M., et al., *Development and evaluation of a simulation-based continuing medical education course: beyond lectures and credit hours.* American journal of surgery, 2015. **210**(4): p. 603-9.
98. Vassiliou, M.C., et al., *FLS and FES: comprehensive models of training and assessment.* Surgical Clinics of North America, 2010. **90**(3): p. 535-558.
99. Fried, G.M., et al., *Proving the value of simulation in laparoscopic surgery.* Annals of Surgery, 2004. **240**(3): p. 518-528.
100. Scott, D.J., et al., *Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training.* Surgical endoscopy, 2008. **22**(8): p. 1887-93.
101. Fraser, S.A., et al., *Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system.* Surg Endosc, 2003. **17**(6): p. 964-7.
102. Okrainec, A., et al., *Trends and results of the first 5 years of Fundamentals of Laparoscopic Surgery (FLS) certification testing.* Surgical endoscopy, 2011. **25**(4): p. 1192-8.
103. Swanstrom, L.L., et al., *Beta test results of a new system assessing competence in laparoscopic surgery.* Journal of the American College of Surgeons, 2006. **202**: p. 62-69.
104. Borman, K.R., et al., *Changing demographics of residents choosing fellowships: longterm data from the American Board of Surgery.* J Am Coll Surg, 2008. **206**(5): p. 782-8; discussion 788-9.
105. Qureshi, A., et al., *MIS training in Canada: a national survey of general surgery residents.* Surgical endoscopy, 2011. **25**(9): p. 3057-65.
106. Park, A., et al., *Minimally invasive surgery: the evolution of fellowship.* Surgery, 2007. **142**(4): p. 505-11; discussion 511-3.
107. Nepomnayshy, D., et al., *Evaluation of advanced laparoscopic skills tasks for validity evidence.* Surg Endosc, 2015. **29**(2): p. 349-54.
108. Dehabadi, M., B. Fernando, and P. Berlingieri, *The use of simulation in the acquisition of laparoscopic suturing skills.* International journal of surgery (London, England), 2014. **12**(4): p. 258-68.
109. Aggarwal, R., et al., *Skills acquisition for laparoscopic gastric bypass in the training laboratory: an innovative approach.* Obes Surg, 2007. **17**(1): p. 19-27.
110. Watanabe, Y., et al., *New models for advanced laparoscopic suturing: taking it to the next level.* Surgical Endoscopy and Other Interventional Techniques, 2016. **30**(2): p. 581-587.
111. Kolozsvari, N.O., et al., *Sim one, do one, teach one: considerations in designing training curricula for surgical simulation.* J Surg Educ, 2011. **68**(5): p. 421-7.
112. Stefanidis, D., *Optimal acquisition and assessment of proficiency on simulators in surgery.* Surg Clin North Am, 2010. **90**(3): p. 475-89.

113. Scott, D.J., et al., *New directions in simulation-based surgical education and training: validation and transfer of surgical skills, use of nonsurgeons as faculty, use of simulation to screen and select surgery residents, and long-term follow-up of learners.* Surgery, 2011. **149**(6): p. 735-44.
114. Watanabe, Y., et al., *Camera navigation and cannulation: validity evidence for new educational tasks to complement the Fundamentals of Laparoscopic Surgery program.* Surgical endoscopy, 2015. **29**(3): p. 552-7.
115. Stefanidis, D., C.E. Acker, and F.L. Greene, *Performance goals on simulators boost resident motivation and skills laboratory attendance.* Journal of surgical education, 2010. **67**(2): p. 66-70.
116. Madan, A.K., et al., *Goal-directed laparoscopic training leads to better laparoscopic skill acquisition.* Surgery, 2008. **144**(2): p. 345-50.
117. Stoller, J., et al., *Are There Detrimental Effects From Proficiency-Based Training in Fundamentals of Laparoscopic Surgery Among Novices? An Exploration of Goal Theory.* Journal of surgical education, 2016. **73**(2): p. 215-21.
118. Derossis, A.M., M. Antoniuk, and G.M. Fried, *Evaluation of laparoscopic skills: a 2-year follow-up during residency training.* Can J Surg 1999. **42**: p. 293-6.
119. Stefanidis, D., et al. *Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: a randomized controlled trial*. Annals of surgery, 2012. **255**, 30-7 DOI: 10.1097/SLA.0b013e318220ef31.
120. Gallagher, A.G., et al., *Virtual Reality Simulation for the Operating Room.* Annals of Surgery, 2005. **241**(2): p. 364-372.
121. Fried, G., et al. *Proving the value of simulation in laparoscopic surgery*. Annals of surgery, 2004. **240**, 518-25; discussion 525-8.
122. Bilgic, E., et al., *Multicenter proficiency benchmarks for advanced laparoscopic suturing tasks.* American journal of surgery, 2016.
123. Kurenov, S., et al., *Development and initial validation of a virtual reality haptically augmented surgical knot-tying trainer for the Autosuture ENDOSTITCH instrument.pdf.* Studies in Health Technology & Informatics, 2009. **142**: p. 145-147.
124. Rinewalt, D., H. Du, and J.M. Velasco, *Evaluation of a novel laparoscopic simulation laboratory curriculum.* Surgery, 2012. **152**(4): p. 550-4; discussion 554-6.
125. Vassiliou, M.C., et al., *A global assessment tool for evaluation of intraoperative laparoscopic skills.* American journal of surgery, 2005. **190**(1): p. 107-13.
126. McKendy, K.M., et al., *Establishing meaningful benchmarks: the development of a formative feedback tool for advanced laparoscopic suturing.* Surg Endosc, 2017.
127. Munshi, A., et al., *Use of a Formative Feedback Tool in Place of an Expert Coach in Laparoscopic Suturing Training: A Randomized Non-inferiority Trial.* Surgical Endoscopy and Other Interventional Techniques, 2016.
128. Peyre, S.E., et al., *Laparoscopic Nissen fundoplication assessment: task analysis as a model for the development of a procedural checklist.* Surgical endoscopy, 2009. **23**(6): p. 1227-32.
129. The American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999) Standard for educational and psychological testing. American Educational Research Associa- tion, Washington
130. Watanabe, Y., et al., *A systematic review of performance assessment tools for laparoscopic cholecystectomy.* Surg Endosc, 2016. **30**(3): p. 832-44.

131. Adrales, G.L., et al., *A valid method of laparoscopic simulation training and competence assessment.* Journal of Surgical Research, 2003. **114**: p. 156-162.
132. Aggarwal, R., et al., *Skills acquisition for laparoscopic gastric bypass in the training laboratory - An innovative approach.* Obesity Surgery, 2007. **17**: p. 19-27.
133. Aggarwal, R., et al., *Training junior operative residents in laparoscopic suturing skills is feasible and efficacious.* Surgery, 2006. **139**: p. 729-734.
134. Barussaud, M.-L., et al., *French intensive training course in laparoscopic surgery (HUGOFirst) on live porcine models: Validation of a performance assessment scale and residents' satisfaction in a prospective study.* Journal of Visceral Surgery, 2015. **153**: p. 15-19.
135. Bilgic, E., et al., *Reliable assessment of operative performance.* American Journal of Surgery, 2016. **211**: p. 426-430.
136. Bingener, J., et al., *Randomized double-blinded trial investigating the impact of a curriculum focused on error recognition on laparoscopic suturing training.* American Journal of Surgery, 2008. **195**: p. 179-182.
137. Birkmeyer, J.D., et al., *Surgical skill and complication rates after bariatric surgery.* New England Journal of Medicine, 2013. **369**: p. 1434-42.
138. Bonrath, E.M., et al., *Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room.* Annals of Surgery, 2015. **262**: p. 1.
139. Bonrath, E.M., et al., *Error rating tool to identify and analyse technical errors and events in laparoscopic surgery.* British Journal of Surgery, 2013. **100**: p. 1080-1088.
140. Botden, S.M.B.I., I.H.J.T. De Hingh, and J.J. Jakimowicz, *Meaningful assessment method for laparoscopic suturing training in augmented reality.* Surgical Endoscopy and Other Interventional Techniques, 2009. **23**: p. 2221-2228.
141. Boyd, T., et al., *Music experience influences laparoscopic skills performance.* JSLS : Journal of the Society of Laparoendoscopic Surgeons / Society of Laparoendoscopic Surgeons, 2008. **12**: p. 292-294.
142. Boza, C., et al., *Simulation-trained junior residents perform better than general surgeons on advanced laparoscopic cases.* Surg Endosc, 2016.
143. Boza, C., et al., *A cadaveric porcine model for assessment in laparoscopic bariatric surgery--a validation study.* Obes Surg, 2013. **23**: p. 589-593.
144. Broe, D., et al., *Construct validation of a novel hybrid surgical simulator.* Surgical endoscopy, 2006. **20**: p. 900-904.
145. Chang, O.H., et al., *Developing an objective structured assessment of technical skills for laparoscopic suturing and intracorporeal knot tying.* Journal of Surgical Education, 2016. **73**: p. 258-263.
146. Crochet, P., et al., *Development of an evidence-based training program for laparoscopic hysterectomy on a virtual reality simulator.* Surg Endosc, 2016.
147. Dath, D., et al., *Toward reliable operative assessment: The reliability and feasibility of videotaped assessment of laparoscopic technical skills.* Surgical Endoscopy and Other Interventional Techniques, 2004. **18**: p. 1800-1804.
148. Dayan, A.B., et al., *A Simple, Low-Cost Platform for Basic Laparoscopic Skills Training.* Surgical Innovation, 2008. **15**: p. 136-142.
149. Doyle, J.D., E.M. Webber, and R.S. Sidhu, *A universal global rating scale for the evaluation of technical skills in the operating room.* American Journal of Surgery, 2007. **193**: p. 551-555.

150. Egi, H., et al., *Evaluating the correlation between the HUESAD and OSATS scores: Concurrent validity study.* Minimally invasive therapy & allied technologies : MITAT : official journal of the Society for Minimally Invasive Therapy, 2012. **5706**: p. 1-6.
151. Enciso, S., et al., *Validation of a structured intensive laparoscopic course for basic and advanced gynecologic skills training.* International Journal of Gynecology and Obstetrics, 2016. **133**: p. 241-244.
152. Enciso, S., et al., *Validation of a model of intensive training in digestive laparoscopic surgery.* Cirugia espanola, 2016. **94**: p. 70-76.
153. Figert, P.L., et al., *Transfer of training in acquiring laparoscopic skills.* Journal of the American College of Surgeons, 2001. **193**: p. 533-537.
154. Ghaderi, I., et al., *Quantitative and qualitative analysis of performance during advanced laparoscopic fellowship: A curriculum based on structured assessment and feedback.* American Journal of Surgery, 2015. **209**: p. 71-78.
155. Goff, B., et al., *Assessment of resident surgical skills: Is testing feasible?* American Journal of Obstetrics and Gynecology, 2005. **192**: p. 1331-1340.
156. Goff, B.A., et al., *Development of a bench station objective structured assessment of technical skills.* Obstetrics and Gynecology, 2001. **98**: p. 412-416.
157. Goff, B.A., et al., *Development of an objective structured assessment of technical skills for obstetric and gynecology residents.* Obstetrics and Gynecology, 2000. **96**: p. 146-150.
158. Goff, B.A., et al., *Surgical skills assessment: A blinded examination of obstetrics and gynecology residents.* American Journal of Obstetrics and Gynecology, 2002. **186**: p. 613-617.
159. Hiemstra, E., et al., *Value of an objective assessment tool in the operating room.* Canadian Journal of Surgery, 2011. **54**: p. 116-122.
160. Husslein, H., et al., *The Generic Error Rating Tool: A Novel Approach to Assessment of Performance and Surgical Education in Gynecologic Laparoscopy.* Journal of Surgical Education, 2015. **72**: p. 1259-1265.
161. King, C.R., et al., *Development and Validation of a Laparoscopic Simulation Model for Suturing the Vaginal Cuff.* Obstetrics and gynecology, 2015. **126 Suppl**: p. 27S-35S.
162. Kowalewski, K.-F., et al., *Development and validation of a sensor- and expert model-based training system for laparoscopic surgery: the iSurgeon.* Surgical Endoscopy, 2016.
163. Kowalewski, T.M., et al., *Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks.* Journal of Urology, 2016. **195**: p. 1859-1865.
164. Kowalewski, T.M., et al., *Validation of the AUA BLUS Tasks.* Journal of Urology, 2016. **195**: p. 998-1005.
165. Kowalewski, T.M., et al., *Beyond task time: Automated measurement augments fundamentals of laparoscopic skills methodology.* Journal of Surgical Research, 2014. **192**: p. 329-338.
166. Kroeze, S.G.C., et al., *Assessment of laparoscopic suturing skills of urology residents: a pan-European study.* Eur Urol, 2009. **56**: p. 865-872.
167. Kurashima, Y., et al., *A tool for training and evaluation of laparoscopic inguinal hernia repair: The global operative assessment of laparoscopic skills-groin hernia (GOALS-GH).* American Journal of Surgery, 2011. **201**: p. 54-61.
168. Larson, J.L., et al., *Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents.* Surgery, 2005. **138**: p. 640-649.

169. Lee, J.Y., et al., *Basic Laparoscopic Skills Assessment Study: Validation and Standard Setting among Canadian Urology Trainees.* The Journal of urology, 2016.
170. Lee, J.Y., et al., *Laparoscopic warm-up exercises improve performance of senior-level trainees during laparoscopic renal surgery.* Journal of Endourology, 2012. **26**: p. 545-550.
171. Matos-Azevedo, A.M., et al., *Comparison of single access devices during cut and suturing tasks on simulator.* Journal of Surgical Research, 2014. **192**: p. 356-367.
172. Moorthy, K., et al., *Bimodal assessment of laparoscopic suturing skills: Construct and concurrent validity.* Surgical Endoscopy and Other Interventional Techniques, 2004. **18**: p. 1608-1612.
173. Munz, Y., et al., *Curriculum-based solo virtual reality training for laparoscopic intracorporeal knot tying: objective assessment of the transfer of skill from virtual reality to reality.* American Journal of Surgery, 2007. **193**: p. 774-783.
174. 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.* Surgery Today, 2013. **43**: p. 271-275.
175. Peyré, S.E., et al., *Reliability of a procedural checklist as a high-stakes measurement of advanced technical skill.* American Journal of Surgery, 2010. **199**: p. 110-114.
176. Peyre, S.E., et al., *Laparoscopic Nissen fundoplication assessment: task analysis as a model for the development of a procedural checklist.* Surgical endoscopy, 2009. **23**: p. 1227-32.
177. Poudel, S., et al., *Development and validation of a checklist for assessing recorded performance of laparoscopic inguinal hernia repair.* American Journal of Surgery, 2015. **212**: p. 468-474.
178. Schijven, M.P., et al., *Transatlantic comparison of the competence of surgeons at the start of their professional career.* British Journal of Surgery, 2010. **97**: p. 443-449.
179. Sharma, M., D. Macafee, and A.F. Horgan, *Basic laparoscopic skills training using fresh frozen cadaver: A randomized controlled trial.* American Journal of Surgery, 2013. **206**: p. 23-31.
180. Stelzer, M.K., et al., *Dry Lab Practice Leads to Improved Laparoscopic Performance in the Operating Room.* Journal of Surgical Research, 2009. **154**: p. 163-166.
181. Takazawa, S., et al., *Video-Based Skill Assessment of Endoscopic Suturing in a Pediatric Chest Model and a Box Trainer.* Journal of Laparoendoscopic & Advanced Surgical Techniques, 2015. **25**: p. 445-453.
182. Tjiam, I.M., et al., *Program for laparoscopic urological skills assessment: setting certification standards for residents.* Minimally invasive therapy & allied technologies : MITAT : official journal of the Society for Minimally Invasive Therapy, 2013. **22**: p. 26-32.
183. Tunc, L., et al., *Evaluation of applied laparoscopic urology course using validated checklist.* JSLS, 2013. **17**: p. 300-305.
184. Tunitsky-Bitton, E., K. Propst, and T. Muffly, *Development and validation of a laparoscopic hysterectomy cuff closure simulation model for surgical training.* American journal of obstetrics and gynecology, 2016. **214**: p. 392.e1-392.e6.
185. Van Sickle, K.R., et al., *Construct validity of an objective assessment method for laparoscopic intracorporeal suturing and knot tying.* American Journal of Surgery, 2008. **196**: p. 74-80.

186. Van Sickle, K.R., et al., *Prospective, Randomized, Double-Blind Trial of Curriculum-Based Training for Intracorporeal Suturing and Knot Tying.* Journal of the American College of Surgeons, 2008. **207**: p. 560-568.
187. Van Sickle, K.R., E.M. Ritter, and C.D. Smith, *The pretrained novice: using simulation-based training to improve learning in the operating room.* Surgical innovation, 2006. **13**: p. 198-204.
188. Varas, J., et al., *Significant transfer of surgical skills obtained with an advanced laparoscopic training program to a laparoscopic jejunojejunostomy in a live porcine model: Feasibility of learning advanced laparoscopy in a general surgery.* Surgical Endoscopy and Other Interventional Techniques, 2012. **26**: p. 3486-3494.
189. Watanabe, Y., et al., *Psychometric properties of the Global Operative Assessment of Laparoscopic Skills (GOALS) using item response theory.* The American Journal of Surgery, 2016: p. 2-5.
190. Weizman, N.F., et al., *Design and validation of a novel assessment tool for laparoscopic suturing of the vaginal cuff during hysterectomy.* Journal of Surgical Education, 2014. **72**: p. 212-219.
191. Yeung, C., et al., *Video assessment of laparoscopic skills by novices and experts: implications for surgical education.* Surgical endoscopy, 2017.
192. Zhao, Z., et al., *Intensive laparoscopic training shortens the learning curve of laparoscopic suturing in surgical postgraduate students: feasible or not?* Journal of endourology / Endourological Society, 2012. **26**: p. 895-902.
193. Reznick, R., et al., *Testing technical skill via an innovative "bench station" examination.* Am J Surg, 1997. **173**(3): p. 226-30.
194. *The general surgery milestone project.* J Grad Med Educ, 2014. **6**(1 Suppl 1): p. 320-8.
195. Crossley, J., et al., *Generalisability: A Key to Unlock Professional Assessment.* Medical Education 2002. **36**: p. 972-978.
196. Cronbach, L.J., N. Rajaratnam, and G.C. Gleser, *Theory of generalizability: A liberation of reliability theory.* The British Journal of Statistical Psychology, 1963. **16**: p. 137-163.
197. Kramp, K.H., et al., *Validity and reliability of global operative assessment of laparoscopic skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy.* Journal of surgical education, 2015. **72**(2): p. 351-8.
198. Vassiliou, M.C., et al., *Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances.* Surgical innovation, 2007. **14**(3): p. 211-6.
199. Shavelson, R.J. and N.M. Webb, *Generalizability Theory: A Primer*. 1991, Newbury Park, CA: Sage Publications.
200. Brennan, R.L., *Generalizability Theory*. 2001, New York: Springer.
201. Gumbs, A.A., N.J. Hogle, and D.L. Fowler, *Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills.* Journal of the American College of Surgeons, 2007. **204**(2): p. 308-13.
202. Williams, R.G., et al., *A template for reliable assessment of resident operative performance: assessment intervals, numbers of cases and raters.* Surgery, 2012. **152**(4): p. 517-24; discussion 524-7.
203. Williams, R.G., et al., *A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study.* Annals of surgery, 2012. **256**(1): p. 177-87.

204. Beckman, T.J., D.A. Cook, and J.N. Mandrekar, *What is the validity evidence for assessments of clinical teaching?* J Gen Intern Med, 2005. **20**(12): p. 1159-64.
205. Castellvi, A.O., et al., *Maintaining proficiency after fundamentals of laparoscopic surgery training: a 1-year analysis of skill retention for surgery residents.* Surgery, 2009. **146**(2): p. 387-93.

CHAPTER 7: APPENDICES

Appendix 2.2.1: Search strategy for MEDLINE

1 exp Laparoscopy/ (90382)
2 Suture Techniques/
(43044) 3 Sutures/
(16421)
4 2 or 3 (54604)
5 1 and 4 (2898)
6 ((extracorporeal or extra-corporeal or intracorporeal or intra-corporeal or laparo*) adj3 (sutur* or knot*)).tw,kf. (2117)
7 5 or 6 (4242)
8 ed.fs. (261740)
9 education, medical, graduate/ (26742)
10 "internship and residency"/ (44310)
11 exp Teaching/ (80605)
12 exp Learning/ (362731)
13 (curricul* or educat* or instruct* or learn* or teach* or train* or
tuition).tw,kf. (1305770) 14 or/8-13 (1650389)
15 7 and 14 (747)
16 Limit 15 to English (696)
17 remove duplicates from 16 (653)
18 limit 17 to yr="2000 Current" (607)

Appendix 2.2.2: Studies reporting data on development and/or validation of simulation tasks for laparoscopic suturing (stars in the first column represent the number of validity sources addressed (1-5 stars); details can be found in Appendix 2.2.3)

Author /Year	Platform		Tasks	Metrics	H/D	Warm-up/ teaching	Results	Extra notes
Leeds S/2017 *	Box trainer	Synthetic suturing pad	ICK (continuous suturing)	1) Time 2) Dots on target (DoT) 3) Total deviation (D) 4) Number of attempts to reach proficiency	+/+	+	Experts had significantly better time and took less attempts to reach proficiency. DoT and D was better for experts, but not significantly	Participants completed the task using Endo360 and traditional laparoscopic technique. The results apply to both devices.
Kowalewski KF/2017 **	Box trainer	Silicone suture pad	ICK (4 sutures)	1) Motion analysis (author's new system called iSurgeon, includes multiple parameters) 2) OSATS (Chang et al, generic and LS specific items)	+/	+	iSurgeon system and OSATS can distinguish between experts and novices. The 2 metrics had strong correlation	
Rosser JC/2017 *	Box trainer	Porcine intestine	ICK (number of sutures NS)	Time and error	+/	NS	ICK task correlates with case experience and Monkey Ball videogame	They have 2 video games (Monkey Ball and Underground), FLS peg transfer, pea drop, and ICK. Result for each task is separate.
Sleiman Z/2017 **	1) Box trainer (homemade) 2) Box trainer (standard)	1,2) Synthetic suturing pad	1,2) ICK (1 suture)	1,2) Stitch out of the dot, tear in the tissue	+/	+	The metrics can distinguish between experts and novices. Participants found value in homemade	They have 4 tasks; 1 being ICK. Result for each task is separate

							trainer for home training, and value in both trainers for training in general	
Yeung C/2017 *	Box trainer	Penrose	ICK (1 suture)	1) Overall performance (Likert scale 1-5) 2) Number of times raters changed the score	+/	NS	Consistency between expert and non-expert raters were high. Experts changed their scores fewer times (not reach statistical significance)	
Deal SB/2016 *	Box trainer	Penrose	FLS ICK (1 suture)	GOALS 20pt	+/	NS	Assessment using crowd-source works	
Kowalewski KF/2016 **	Box trainer	Silicone suture pad	ICK (4 sutures)	1) Motion analysis (author's new system called iSurgeon) 2) LSIKT-GRS + LSIKT-CL	+/	+	iSurgeon system is a valid way to assess skill	
Sanchez-Margallo JA/2016 **	Box trainer	Porcine stomach	ICK (3 sutures)	1) Moorthy CL 27pt 2) Motion analysis (Micron Tracker)	+/	NS	Validity evidence was gathered for assessment using motion metrics, but need methods to assess quality of the suturing.	
Poursartip B/2016 *	Box trainer	Synthetic skin model	- pass needle through incision - ICK (1 suture)	Motion analysis (SIMIS system; metrics are potential energy, kinetic energy, work)	+/	NS	Trainees with more laparoscopic case experience scored better	Participants performed the tasks 4 times. Their 3rd trial scores were used
Zdichavsky M/2016 *	Box trainer	Porcine stomach and intestine (gastro-jejunostomy)	ICK (continuous suturing)	1) Time 2) Accuracy 3) Anastomotic width 4) Pressure resistance	+/	+	Trainees with more laparoscopic case experience scored better	The needle driver is a new steerable instrument called r2. It has active

								tip deflection and tip and shaft rotation.
Veneziano D/2016 **	Box trainer	1) Penrose 2) Silicone based SimPORTAL MVI model	1) ICK (1 suture) 2) They could use any technique to stop the bleeding	1) Time and error (FLS scoring) 2) Blood loss, number of stitches	+/	+	Trainees in higher training levels scored better (statistical significance not reported). Trainees agreed that MVI model can assess repair of MVI skill, is realistic, and should be included in the curriculum	The MVI model was perfused with synthetic blood. The room had low lights and pulse sounds to simulate OR.
Buckley CE/2015 *	Box trainer (ProMIS augmented reality)	Synthetic suturing skin	ICK (1 suture)	1) Computer generated metrics in the simulator 2) OSATS 25pt 3) Time and error(FLS scoring) 4) Zone calculations from a new software	+/	+	The zone calculation is a good metric for this task	
Chang OH/2015 **	Box trainer	Penrose	FLS ICK (1 suture)	1) LSIKT-GRS 2) LSIKT-CL	+/	NS	The assessment tools are able to measure suturing skills	
Stefanidis D/2015 *	Box trainer	Penrose	FLS ICK (1 suture)	Time and error (FLS scoring)	+/	+	PGY level and case experience were good predictors of simulation performance.	They have 3 laparoscopic (1 is FLS ICK) and 5 open tasks. Result for each task is separate
Trudeau MO/2015 *	Box trainer	Penrose	ICK (1 suture)	Motion analysis (velocity, acceleration, range)	+/	NS	Trainees with more laparoscopic case experience scored better	

							(range was not statistically different)	
Uemura M/2015 *	Box trainer (Augmented reality)	Synthetic intestinal model	ICK (3 sutures)	1) Time 2) Air pressure leak 3) Number of full-thickness sutures 4) Suture tension 5) Wound-opening area	+/	+	The metrics can distinguish between experts and novices. Expert benchmarks were developed	The simulation platform is called Suture Simulator Instruction Evaluation Unit.
Kowalewski TM/2014 **	Box trainer Simulab EDGE platform	Penrose	FLS ICK (1 suture)	1) Time and error 2) p-OSATS 3) Motion analysis (automatically calculated by the EDGE platform)	+/	NS	The various metrics used had good correlation with each other, and higher training levels/case experience scored better, and p-OSATS showed good inter-rater reliability.	
Trejos AL/2014 *	Box trainer	Foam and silicone	- Needle driving (no knot-tying) - ICK (1 suture)	Motion analysis (SIMIS system; metrics based on time, position, and force)	+/	NS	Force-based metrics are better able to differentiate experts from novices	The task is a procedure that has 5 subtasks that are assessed separetly; 1 driving needle through tissue, 1 ICK. Result for each task is separate. Participants completed the procedure 4 times
Stefanidis D/2014 *	Box trainer	Penrose	FLS ICK (1 suture)	Time and error (FLS scoring)	+/	+	Validity evidence was gathered for the task	They have 3 laparoscopic (1 is FLS ICK) and 5 open tasks. Result for each task is

								separate. Participants completed each task 2 times
Lusch A/2014 *	Box trainer	1) Rings 2) Silicone suture slab 3) Suture slab	1)Passing needle through ring without knot-tying 2) ICK (1 suture) 3) ICK (multiple sutures)	Quantity score: 1) Number of passed rings 2) Number of ties within knot 3) Number of suture throws Quality score: 1) Number of missed attempts 2,3) Distance between knot and suture	+/	NS	Validity evidence was gathered for the 2D and 3D systems	The study was comparing 2D vs 3D systems. So the validity evidence is for both. They have 6 tasks, 3 suturing related. Participants completed all tasks in 2D and 3D. Result for each task and system is separate.
Escamirosa F/2014 ***	Box trainer (EndoVi S augment ed reality training system)	Penrose	ICK (1 suture)	Metrics of the simulator	+/	NS	Validity evidence was gathered for the task	They have 4 tasks; 1 being ICK. Result for each task is separate
Romero P/2014 *	Box trainer	Synthetic suturing pad	ICK (3 sutures)	1) Time 2) Knot quality (5pt scale) 3) Accuracy 4) Moorthy CL 23 pt	+/	+	Experts had better scores than junior trainees	
Egi H/2013 **	Box trainer	-	ICK (1 suture)	1) Moorthy CL 29pt 2) OSATS Buckley 2015	+/	NS	Good correlation between the LS task scores and HUESAD(this is an augmented reality simulator & doesn't assess LS) motion analysis task scores	

Boza C/2013 *	Box trainer	Porcine intestine (Jejuno-jejunostomy)	- ICK (1 suture) - ICK (continuous suturing)	1) Motion analysis (ICSAD (time, path length, total no of movements)) 2) ALRYGB	+/	NS	Strong correlation between performance in the simulator and the OR (completed jejuno-jejunostomy of a LRYGBP and assessments were identical to simulator)	
Hennessey I/2013 **	1) Box trainer(FLS) 2) Box trainer (eoSim)	1) Penrose 2) NS(looks like a fabric)	1) ICK (1 suture) 2) ICK (1 suture)	1) Time and error (FLS scoring) 2) NS	+/	+	Validity evidence was gathered for the 2 platforms	There were 3 tasks: object transfer, cutting, ICK. Result for each task is separate. Participants completed both platforms, but the order of which one they start with was randomly selected.
Bahsoun AN/2013 *	Box trainer	Hoops	Passing needle through 3 hoops without knot-tying	-	-	NS	Experts found that the trainer has high training capacity and performance (video, light etc)	The box is made of cardboard, and use iPad2 as camera and monitor.
Palter V/2012 *	VR(LAPSIM)	-	ICK (1 suture)	Metrics of the simulator(time, path length, angular path)	+/	+	Expert benchmarks for the tasks were developed	First, they did a Delphi study to determine which tasks on the VR simulator should be included in the proficiency-based VR technical skills curriculum

								for colorectal surgery. Then, experts completed the tasks to determine benchmarks (8 tasks are included in the final curriculum, 1 suturing. Result for each task is separate).
Pagador JB/2012 *	Box trainer	Carcass stomach	FLS ICK (1 suture)	Augmented reality haptic (ARH, motion analysis)	+/	NS	Different metrics showed significant difference between the levels for the 4 subtasks.	LS is decompsoed into 4 subtasks; needle puncture, first knot, second knot, third knot.
Sharma M/2012 *	Fresh Frozen Cadaver (FFC)	Mesenteric rifts	-ICK (1 suture) -ECK (1 suture)	GOALS 20pt	+/	+	The metrics can distinguish between experts and novices. High inter-rater reliability for ICK task. Expert benchmarks were developed	They have 5 tasks, including ICK and ECK. Result for each task is separate.
Strickland A/2011 *	Box trainer (ProMIS augmented reality)	Lamb liver	ICK (2 sutures (one at an easier and one at a more difficult location))	Metrics of the simulator(time, path length)	+/	+	The metrics can distinguish between experts and novices	They have 4 tasks; 1 easy stitch, 1 hard stitch. Results for each task is separate
Hiemstra E/2011 *	Box trainer	Synthetic suturing pad	ICK (1 suture, they performed the task 3 times consecutively)	Motion analysis (TrEndo; time, path length, motion in depth, motion smoothess)	+/	+	Experts had better scores than residents and medical students	LS skills improved from 1st to 3rd trial for medical students and residents, but experts were consistent.

Kobayashi SA/2011 **	Box trainer	Penrose	-FLS ICK (1 suture) -FLS ECK (1 suture)	Time and error (FLS scoring)	+/	+	The metrics can distinguish between experts and novices. Experts found the simulator realistic and could help improve skill	They have 5 tasks, including ICK and ECK. Result for each task is separate
Zheng B/2010 *	Box trainer	Synthetic soft tissue	ICK (interrupted, as many sutures as possible in 6min) while also responding to a visual cue	1) Time and error (FLS scoring) 2) Number of sutures completed 3) Error scoring for the visual detection task (secondary)	+/	NS	Experienced surgeons performed more sutures, had higher quality, and scored better in the secondary task	While performing the suturing tasks, participants were asked to respond to visual cues correctly, and authors hypothesize that experts will have more space in their cognitive capacity to perform this secondary task.
Yamaguchi S/2010 *	Box trainer	Synthetic suturing pad	ICK (1 suture)	Motion analysis (AURORA; time, path length, average speed)	+/	+	Experts had better scores than residents	
Horeman T/2010 *	Box trainer	Synthetic skin model	Needle driving (no knot-tying required)	Motion analysis (force)	+/	NS	The force metric can distinguish between experts and novices	Participants completed the task twice.
Botden SM/2009 **	Box trainer (ProMIS augmented reality)	Synthetic suturing pad	ICK (1 suture)	1) Time 2) Knot strength 3) Time spent in correct area 4) Botden 2009 LS specific tool	+/	NS	Using augmented reality is a way to measure LS skills	
Kroeze SGC/2009 **	Box trainer	Synthetic suturing pad	ICK (1 suture)	Moorthy CL 29pt	+/	+	There was a relationship	

							between score and PGY level	
Lin D/2009 **	1) Box trainer(ProMIS augmented reality) 2) VR simulator (Surgical SIM)	1) Latex drain 2) -	1,2) ICK (1 suture)	Metrics of the simulators	+/	+	Validity evidence was gathered for the 2 platforms	
Xerouli s G/2009 *	Box trainer	Penrose	FLS ICK (1 suture)	1) Time and error (FLS scoring) 2) Motion analysis (ICSAD)	+/	+	The metrics can distinguish between experts and novices, and they correlate significantly	They have 4 tasks; 1 being ICK. Result for each task is separate
Dayan AB/2008 **	Box trainer	Synthetic suturing pad	ICK (1 suture)	1) Moorthy CL 27pt 2) Time	+/	+	The simple, low-cost laparoscopic training platform has good validity evidence	There were 3 tasks: rope passing, peg transfer, and intracorporeal knot tying). Participants completed all 3 tasks, but results separate for each. Short warm-up session. Max time 15 min for LS
Oostema JA/2008 *	Box trainer (ProMIS augmented reality)	NS	ICK (1 suture)	Metrics of the simulator (time, path length, smoothness)	+/	+	Participants with more laparoscopic case experience scored better	They have 4 tasks; 1 being ICK. They performed 3 repetitions. Result for each task is separate
Zheng B/2007 **	Box trainer	Penrose	ICK (1 suture)	1) Time and accuracy (LISETT score) 2) Self-rated team quality scores	+/	NS	Validity evidence was gathered for the LISETT system that is meant to enhance team skills	They have 2 tasks; peg transfer and ICK. The results are for the combination score of the 2 tasks

Stefanidis D/2007 *	Box trainer	Penrose	ICK (interrupted, as many sutures as possible in 10min) while also responding to a visual cue	1) Time and error (FLS scoring) 2) Number of sutures completed 3) Error scoring for the secondary task	+/	NS	Experienced surgeons performed more sutures, had higher quality, and scored better in the secondary task.	While performing the suturing tasks, participants were also asked to complete a visual-spacial secondary task for attention (2 tasks performed simultaneously).
Botden SM/2007 **	1) Box trainer (ProMIS augmented reality) 2) VR(LAPSIM)	1) Synthetic suturing pad 2) -	1) ICK (1 suture) 2) ICK (1 suture)	Metrics of the simulators	+/	+	Participants found augmented reality (AR) to be more realistic and had better training quality. 4/5 AR metrics and 1/5 VR metrics can distinguish experts and novices.	Participants performed 2 tasks (1 LS) in each simulator. Result for each task and simulator is separate
Broe D/2006 **	Box trainer (ProMIS augmented reality)	4 hoops	Passing needle through 4 hoops without knot-tying	OSATS 25pt	-	+	Senior residents scored better than junior, and there was good inter-rater reliability.	There were 3 tasks: Laparoscopic orientation, dissection and suturing. The results are for the combination of 3. They also had a CL for each of the 3 tasks, however, we do not know what the items are, so we only looked at GRS

Dubrowski A/2006 *	Box trainer	Synthetic skin model	- (no knot-tying, only passing needle from left part of the tissue to right)	Motion analysis	-	+	Most motion analysis metrics were sensitive to residency level	They performed suturing 10 times, without knot-tying
Van Sickle KR/2005 *	Box trainer (ProMIS augmented reality)	Latex glove finger	- (Passing needle through 5 paired circles without knot-tying)	Metrics of the simulator (time, path length, smoothness)	-	+	The metrics can distinguish between experts and novices	
Duffy AJ/2005 *	VR(LAPSIM)	-	ICK (1 suture)	Metrics of the simulator	+/	+	The VR metrics can distinguish between experts and novices	They have 8 tasks, 1 being ICK. Result for each task is separate
Moorthy K/2004 *	Box trainer	Synthetic suturing pad	ICK (2-3 sutures)	1) Moorthy CL 29pt 2) Motion analysis (ICSAD (time, path length))	+/	+	The various metrics used had good correlation with each other	They had experts and novices. Novices received video-based inctructions before completing the task.
Figert PL/2001 *	Box trainer	Organ-shaped foam rubber	ICK (3 sutures) for 3 different knot techniques	1) Time 2) 14 item error rating-tool(LS-ERS)	+/	+	Trainees with more laparoscopic case experience scored better	They completed a pre-assessment session 1) Didactic session(instruments, trocar and camera placement, knot-tying techniques) 2) Demonstration & written instructions on the techniques
Keyser EJ/2000 *	Box trainer (2 different	Penrose	-FLS ICK (1 suture) -FLS	Time and error (FLS scoring)	+/	+	Trainees with more laparoscopic	They have 7 tasks, including

	boxes were used)	ECK (1 suture)		case experience scored better in both boxes, and there was moderate correlation between the 2 box scores	ICK and ECK. Result for each task is separate

H/D Handsewn/Device-assisted

LS Laparoscopic suturing, ICK Intracorporeal knot-tying, FLS Fundamentals of Laparoscopic Surgery, GOALS Global Operative Assessment of Laparoscopic Surgery, CL Checklist, OSATS Objective Structured Assessment of Technical Skills, ERS Error Rating Scale, NS Not Specfied, p-OSATS Psychomotor-OSATS, ICSAD Imperial College Surgical Assessmet Device, ALRYGB Assessment of Laparoscopic Roux en-Y Gastric Bypass, GRS Global Rating Scale, LSIKT Laparoscopic suturing and intracorporeal knot-tying, LISETT Legacy Inanimate System for Endoscopic Team Training, VR Virtual Reality, SIMIS Sensorized instrument-based minimally invasive surgery system, pt Point, ECK Extracorporeal knot-tying, MVI Major vessel injury, OR Operating room.

Appendix 2.2.3: Details regarding the validity evidence of the laparoscopic suturing simulation platforms

Author/Year	Sources of validity targeted*
Leeds S/2017	R(training level)
Kowalewski KF/2017	R(training level; correlation between metrics) I(test-retest reliability for both metrics)
Rosser JC/2017	R(case experience)
Sleiman Z/2017	C(expert opinion) R(training level)
Yeung C/2017	I(IRR, test-retest)
Deal SB/2016	I(IRR, internal consistency)
Kowalewski KF/2016	R(case experience for both metrics; correlation of the metrics) I(test-retest for both metrics)
Sanchez-Margallo JA/2016	R(Case experience; correlation between the metrics) I(IRR for assessment tool)
Poursartip B/2016	R(case experience)
Zdichavsky M/2016	R(case experience)
Veneziano D/2016	C(participant opinion for MVI model) R(training level)
Buckley CE/2015	R(case experience for zone; correlation of the metrics)
Chang OH/2015	R(training level) I(rater reliability)
Stefanidis D/2015	R(training level/case experience)
Trudeau MO/2015	R(case experience)
Uemura M/2015	R(training level; expert benchmarks)
Kowalewski TM/2014	R(training level/case experience; correlation between the metrics) I(only for p-OSATS, rater reliability)
Trejos AL/2014	R(training level)
Stefanidis D/2014	R(training level/case experience)
Lusch A/2014	R(training level)
Escamirosa F/2014	C(expert opinion from a questionnaire) R(training level) I(Internal consistency for 4 tasks together)
Romero P/2014	R(training level)
Egi H/2013	R(correlation between metrics) I(IRR for the assessment tools)
Boza C/2013	R(correlating simulation and OR scores for ICSAD and ALRYGB)
Hennessey I/2013	C(expert opinion) R(training level; correlation between FLS and eoSim ICK)
Bahsoun AN/2013	C(expert opinion)
Palter V/2012	R(expert benchmark)

Pagador JB/2012	R(training level)
Sharma M/2012	R(training level; expert benchmarks) I(IRR for ICK task)
Strickland A/2011	R(training level)
Hiemstra E/2011	R(training level)
Kobayashi	C(expert opinion)
SA/2011	R(training level)
Zheng B/2010	R(training level)
Yamaguchi S/2010	R(training level)
Horeman T/2010	R(training level)
Botden SM/2009	R(training level; correlation between the metrics) I(IRR for assessment tool)
Kroeze SGC/2009	R(training level) I(IRR)
Lin D/2009	C(expert opinion) R(training level)
Xeroulis G/2009	R(training level; correlation between the metrics)
Dayan AB/2008	R(case experience for both metrics) I(IRR)
Oostema JA/2008	R(case experience)
Zheng B/2007	C (expert and participant opinion) R(training level/case experience; correlation between LISETT score and team quality)
Stefanidis D/2007	R(training level)
Botden SM/2007	C(participant opinion for AR) R(training level)
Broe D/2006	R(training level) I(IRR)
Dubrowski A/2006	R(training level)
Van Sickle KR/2005	R(training level)
Duffy AJ/2005	R(training level)
Moorthy K/2004	R(case experience; metrics correlated with each other)
Figert PL/2001	R(case experience)
Keyser EJ/2000	R(case experience, correlation between 2 platforms)

*C Content, RP Response Process, I Internal structure, R Relationship to Other Variables, CO Consequences

IRR Inter-rater reliability, p-OSATS Psychomotor-Objective Structured Assessment of Technical Skills, OR Operating-room, ICSAD Imperial College Surgical Assessmet Device, ALRYGB Assessment of Laparoscopic Roux en-Y Gastric Bypass, ICK Intracorporeal knot-tying, FLS Fundamentals of Laparoscopic Surgery, LISETT Legacy Inanimate System for Endoscopic Team Training

Appendix 3.2.1: Number of residents who took the Fundamentals of Laparoscopic Surgery test (according to each year and PGY Level)

PGY Level	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
1	1	3	5	24	13	12	25	20	3	106
2	5	6	13	45	57	86	55	66	25	358
3	12	18	60	109	155	245	131	177	47	954
4	53	226	271	321	347	418	159	236	48	2079
5	55	208	498	437	548	590	214	308	185	3043
6	23	32	33	26	35	19	14	15	10	207
7	19	50	65	124	100	64	32	26	5	485
Total	168	543	945	1086	1255	1434	630	848	323	7232
PGY post-graduate year, PGY 6: Fellow, PGY7: Attending										

Appendix 4.2.1: MEDLINE Search Strategy

1 *Laparoscopy/
2 exp *Laparoscopy/mt, ed
3 exp *Minimally Invasive Surgical Procedures/ed, is, mt, px, st, sn, td [Education, Instrumentation, Methods, Psychology, Standards, Statistics & Numerical Data, Trends]
4 *specialties, surgical/ or *colorectal surgery/ or *general surgery/ or *gynecology/ or *neurosurgery/ or *obstetrics/ or *orthopedics/ or *otolaryngology/ or *surgery, plastic/ or *thoracic surgery/ or *urology/
5 exp *Surgical Procedures, Operative/ed, is, mt, px, sn, sn, td
6 *Video-Assisted Surgery/ed, is, mt, st, sn, td, ut [Education, Instrumentation, Methods, Standards, Statistics & Numerical Data, Trends, Utilization]
7 exp *General Surgery/ed, mt
8 or/1-7
9 exp *Teaching/ed, mt, st, td [Education, Methods, Standards, Trends]
10 exp *Hospitals, Teaching/ed, mt, og, st, td [Education, Methods, Organization & Administration, Standards, Trends]
11 exp *"Internship and Residency"/ed, mt, og, st, sn, td or (Internship and Residency).mp.
12 *Education, Medical, Continuing/
13 *Education, Medical, Graduate/
14 *Education, Medical, Undergraduate/
15 exp *Curriculum/
16 or/9-14
17 *Competency-Based Education/
18 exp *Professional Competence/
19 (proficiency-based adj3 (train* or educat* or teach* or test* or scor* or assess* or evaluat* or apprais* or measur* or observ* or rating)).tw,kw,kf,ab,ti.
20 (proficiency-based and (train* or educat* or teach* or test* or scor* or assess* or evaluat* or apprais* or measur* or observ* or rating)).tw,kw,kf,ab,ti.
21 exp suturing method/ or *Suture Techniques/ed, is, st, sn or *Sutures/
22 exp operation duration/ or *Operative Time/
23 *Operating Rooms/
24 or/17-23
25 8 and 16 and 24
26 exp *Students, Medical/
27 *Surgeons/
28 (residen* adj3 educat*).tw,kw,ab,ti.
29 (resident* or residency).ab,tw,kw,ti.
30 (student* or residen* or fellow* or expert* or surgeon* or novice*).tw,kw,kf,ab,ti.
31 ((student* or residen* or fellow* or expert* or surgeon* or novice*) adj3 (train* or educat* or teach* or test* or scor* or assess* or evaluat* or apprais* or measur* or observ* or rating)).tw,kw,kf,ab,ti.
32 (intraoperative and (train* or educat* or teach* or test* or scor* or assess* or evaluat* or apprais* or measur* or observ* or rating)).tw,kw,kf,ab,ti.
33 or/26-32
34 ((sutur* or tie* or tying or knot* or intracorporeal knot*) adj3 (skill* or capacity or competenc* or abilit* or technique* or expert* or proficienc* or dexter* or command* or master* or exploit* or effic* or task* or level* or scor* or teach*)).tw,kw,ab,ti.
35 (intraoperative and (skill* or capacity or competenc* or abilit* or technique* or expert* or proficienc* or dexter* or command* or master* or exploit* or effic* or task* or level* or scor* or teach*)).tw,kw,ab,ti.

36 (((skill* or capacity or competenc* or abilit* or technique* or expert* or proficienc* or dexter* or command* or master* or exploit* or effic* or task* or level*) adj3 (assess* or observ* or measur* or rating or valid* or train* or scor*)) and laparoscop*).tw,kw,ab,ti.
37 ((technical not (technical and no*?technical)) and skill*).tw,kw,ab,ti.
38 (laparoscop* adj3 (simulat* or teach* or setting* or laborat* or train*)).tw,kw,ab,ti.
39 (laparoscop* and (simulat* or teach* or setting* or laborat* or train*)).tw,kw,ab,ti.
40 (laparoscop* adj3 (sutur* or tie* or tying or knot*)).tw,kw,ab,ti.
41 (laparoscop* and (sutur* or tie* or tying or knot*)).tw,kw,ab,ti.
42 technical expertise.tw,kw,ab,ti.
43 (intracorporeal and (knot* or knot?tying* or knot?tie*or sutur*)).tw,kw,kf,ab,ti.
44 (intracorporeal adj3 (knot* or knot?tying* or knot?tie*or sutur*)).tw,kw,kf,ab,ti.
45 (knot* or knot?tie* or knot?tying* or sutur*).tw,kw,ab,ti.
46 ((sutur* or laparoscop*) and (train* or skill*)).tw,kw,ab,ti.
47 ((sutur* or laparoscop*) and checklist*).tw,kw,ab,ti.
48 (laparoscop* adj3 (skill* or capacity or competenc* or abilit* or technique* or expert* or proficienc* or dexter* or command* or master* or exploit* or effic* or task* or level*)).tw,kw,ti,ab.
49 (laparoscop* and (skill* or capacity or competenc* or abilit* or technique* or expert* or proficienc* or dexter* or command* or master* or exploit* or effic* or task* or level*)).tw,kw,ab,ti.
50 (operati* adj3 skill*).tw,kw,ab,ti.
51 (performance adj3 (assess* or evaluat*)).tw,kw,kf,ab,ti.
52 (laparoscop* and (sutur* adj3 expert*)).tw,kw,ab,ti.
53 (transferability adj3 skill*).mp,tw,ab,ti.
54 exp *Decision Making/
55 exp *Learning Curve/ or exp *Learning/
56 ((Quantitative or qualitative) adj3 assess*).tw,kw,kf,ab,ti.
57 (observation* adj3 method*).tw,kw,kf,ab,ti.
58 *Patient Simulation/
59 *User-Computer Interface/
60 (assessment or (assess* adj3 tool*)).tw,kw,ab,ti.
61 box trainer.tw,kw,ab,ti.
62 Checklist/is, mt, st, sn, td, ut [Instrumentation, Methods, Standards, Statistics & Numerical Data, Trends, Utilization]
63 direct observation.tw,kw,kf,ab,ti.
64 Educational Measurement/mt
65 exp *"Task Performance and Analysis"/
66 exp *Psychomotor Performance/
67 exp Teaching Materials/is, mt, td, ut
68 exp *Feedback/
69 Grading.tw,kw,kf.
70 human factor* study.tw,kw.
71 Motion metric*.kw,tw.
72 (objective assess* adj3 skill*).kw,tw,ab,ti.
73 *Observation/mt
74 self-appraisal.tw,kw,kf,ab,ti.
75 *Self-assessment/
76 single-blind method/
77 Double-Blind Method/
78 ((laparoscop or surg*) and train*).tw,kw.
79 *"time and motion studies"/
80 *"Surveys and Questionnaires"/
81 (touch perception or touch).tw,kw.

82 tracking.tw,kw,kf.
83 or/34-52
84 25 and 33 and 83
85 or/53-82
86 *Models, Anatomic/ or models, animal/ or *Manikins/
87 (animal* or cadaver* or manikin* or animal model*).tw,kw.
88 86 or 87
89 84 and 88
90 (("VR" or virtual reality) adj3 (laparoscop* and (train* or simulat*))).tw,kw,ab,ti.
91 ((simulat* or video or computer* or virtual or virtual reality) adj3 laparoscop*).tw,kw,ab,ti.
92 *Computer Simulation/ or computer simulation.tw,kw,ab,ti.
93 *Computer-Assisted Instruction/
94 *Videotape Recording/ or video* recording.tw,kw,ab,ti.
95 ("virtual reality" adj3 train*).tw,kw,kf,ab,ti.
96 Simulation.mp,tw,kw.
97 exp *Simulation Training/
98 Simulat* train*.tw,kw.
99 or/90-98
100 84 and 99
101 84 and 88 and 99
102 84 and 85 and 88
103 84 and 85 and 88 and 99
104 84 and 85 and 99
105 exp *Evaluation Studies as Topic/
106 *Validation studies as Topic/
107 exp *case-control studies/
108 Prospective Studies/
109 Retrospective Studies/
110 Feasibility Studies/
111 105 or 106 or 107 or 108 or 109 or 110
112 "Predictive Value of Tests"/
113 Observer Variation/
114 "Reproducibility of Results"/
115 (structured adj3 assessment).mp,tw,kw,ab,ti.
116 Generalizability.mp,tw,kw,ab,ti.
117 Reliability.mp.
118 Statistics, Nonparametric/
119 or/112-118
120 84 and 88 and 111
121 84 and 88 and 119
122 84 and 99 and 111
123 84 and 99 and 119
124 84 and 85 and 88 and 111
125 84 and 85 and 88 and 119
126 84 and 85 and 99 and 111
127 84 and 85 and 99 and 119
128 25 and 83 and 88 and 99
129 25 and 32 and 35 and 88 and 99
130 25 and 32 and 35 and 88
131 35 or 36 or 48 or 49 or 50
132 25 and 33 and 131

133 40 and 53
134
25 and 83
135 85 and 134
136 111 and 135
137 119 and 135
138 136 or 137
139 limit 138 to (abstracts and english language and yr="1993 -Current")

140 limit 139 to "all child (0 to 18 years)"
141 limit 140 to "all adult (19 plus years)"
142 140 not 141
143 139 not 142
144 33 and 143

Appendix 4.2.2: Types of suturing and knot tying that were used with the assessment tools

	Simulation			OR		
	Interrupted/Continuous	IKT/EKT	Handsewn/Device-assisted	Interrupted/Continuous	IKT/EKT	Handsewn/Device-assisted
Global Rating Scale						
OSATS-GRS						
Original	+/	+/	+/	NS	NS	NS
Dath 2004	+/	NS	NS			
Hiemstra 2011				/+	+/	NS
Crochet 2016	/+	+/	NS			
Birkmeyer 2013				NS	NS	NS
Bingener 2008	+/NS	+/	+/	NS/+	+/NS	NS
Broe 2006	+/+	+/	+/			
Kowalewski 2014	+/	+/	+/			
Antosh 2013				+/	/+	+/
GOALS						
Original	+/	+/	+/	+/+	+/NS	NS
Stelzer 2009	+/	+/+	+/			
Lee 2012				+/	+/	+/
GOALS-GH				/+	+/	NS
LSIKT-GRS	+/	+/	+/			
LVG	+/	+/	+/			
GRS-OS	+/+	+/	+/	+/+	+/	+/
ALRYGB	+/+	+/	+/	+/+	+/NS	+/NS
OCRS						
Original	+/	NS	NS	+/	NS	NS
Ghaderi 2015				/+	+/	NS
ASVC	+/	/+	+/			
ASVC	+/	/+	+/			
LS-GRS	+/	+/	+/			
LS-QRS	+/	+/	+/			
LS-AR	+/	+/	+/			
FLP	/+	+/	NS			
BOSATS				NS	NS	NS
GRITS				/+	+/	NS
OPRs				/+	+/	NS
Checklist						
LS Checklist by Moorthy						

Original	+/	+/	+/			
Munz 2007	+/	+/	+/			
Tjiam 2013	+/	+/	+/			
IKT-CL	+/	+/	+/			
LS-CL	/+	+/	+/			
LSIKT-CL	+/	+/	+/			
LNF-CL	+/	NS	NS	+/	NS	NS
TAPP-CL				/+	+/	NS
ASVC	/+	+/	+/			
Error rating scale						
Van Sickle LS error rating						
Original	+/	+/	+/	+/	+/	+/
Takazawa 2015	+/	+/	+/			
LS-ERS	+/	+/	+/			
LIHR-ERS	/+	+/	NS			
GERT				NS/+	+/NS	NS

OR Operating-room, IKT Intracorporeal knot-tying, EKT Extracorporeal knot-tying
NS Not Specified by the papers
OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, GOALS Global Operative Assessment of Laparoscopic Skills, GOALS-GH Groin Hernia, LSIKT LS and intracorporeal knot tying, LVG Laparoscopic Video Grader, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of LRYGB, OCRS Objective Component Rating Scale, ASVC Assessment for suturing of vaginal cuff, LS-QRS LS-Quality rating scale, LS-AR LS in augmented reality simulator, FLP Fundamentals to laparoscopic procedures, BOSATS Bariatric Objective Structured Assessment of Technical Skill, GRITS Global Rating Index for Technical Skills, OPRs Operative Performance Rating System, IKT Intracorporeal knot tying, LNF-CL Laparoscopic nissen fundoplication-Checklist, TAPP-CL Transabdominal Preperitoneal procedure-Checklist, LS-ERS Laparoscopic suturing-Error rating scale, LIHR-ERS Laparoscopic inguinal hernia repair-Error rating scale, GERT Generic Error Rating Tool.

Appendix 4.2.3: Summary of the validity evidence for assessment tools used in the simulation setting

Table 1: Validity evidence for the assessment tools that were used to **specifically assess laparoscopic suturing in simulation setting.** Each evidence category score is out of 3, with a total score of 15.

		Content	Response Process	Internal Structure	Relations to Other Variables	Consequences	Total
OSATS-GRS	Original	1	2	2	3	0	8
	Bingener 2008	1	1	1	1	0	4
	Broe 2006	1	0	1	2	0	4
	Kowalewski 2014	1	0	1	3	0	5
GOALS	Original	1	0	1	3	0	5
	Stelzer 2009	1	1	1	2	0	5
	LSIKT-GRS	2	0	2	2	0	6
	LSIKT-CL	2	0	2	2	0	6
	ASVC[c]	1	0	1	1	0	3
	ASVC[d]	1	0	1	1	2	5
	LS-GRS	1	0	0	1	0	2
	LS-QRS	1	0	0	1	0	2
	LS-AR	1	0	1	2	0	4
	IKT-CL	2	0	1	0	0	3
	LS-CL	2	0	1	0	0	3
	LS-ERS	1	0	0	1	0	2
LS Checklist by Moorthy	Original	1	1	1	3	0	6
	Munz 2007	1	0	1	2	0	4
	Tijam 2013	1	0	3	0	2	6
Van Sickle LS Error Rating	Original	2	2	1	1	0	6
	Takazawa 2015	1	2	1	1	0	5
	ASVC[e]	2	0	1	1	0	4
	LVG	0	0	2	0	0	2

OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, CL Checklist, LS Laparoscopic Suturing, GOALS Global Operative Assessment of Laparoscopic Skills, LSIKT-GRS LS and intracorporeal knot tying-GRS, ASVC Assessment for suturing of vaginal cuff, LS-QRS LS-Quality rating scale, LS-AR LS in augmented reality simulator, IKT-CL Intracorporeal knot tying-CL, LS-ERS LS-Error rating scale, LVG Laparoscopic Video Grader.

[c,d,e]There are three different tools for ASVC. [c]Tunitsky-Bitton 2016, [d]King 2015, [e]Weizman 2014

Table 2: Validity evidence for the assessment tools that were used to **assess procedures that required laparoscopic suturing in simulation setting** (e.g. simulated laparoscopic nissen fundoplication). The procedures are stated in table 2. Each evidence category score is out of 3, with a total score of 15.

		Content	**Response Process**	**Internal Structure**	**Relations to Other Variables**	**Consequences**	**Total**
OSATS-GRS	Original	1	0	2	0	0	3
	Crochet 2016	0	0	1	1	0	2
	Dath 2004	1	0	2	0	0	3
	GRS-OS	1	0	1	1	0	3
	ALRYGB	1	0	1	2	0	4
	OCRS	2	0	2	0	0	4
	LNF-CL	3	0	0	1	0	4
	FLP	1	0	0	1	0	2
	LIHR-ERS	1	0	0	1	0	2

OSATS-GRS Objective Structured Assessment of Technical Skills-Global rating scale, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of LRYGB, OCRS Objective Component Rating Scale, LNF-CL Laparoscopic nissen fundoplication-Checklist, FLP Fundamentals to laparoscopic procedures, LIHR-ERS Laparoscopic inguinal hernia repair-Error rating scale.

Appendix 4.2.4: Details of the validity evidence of each assessment tool

Table 1: Validity evidence for the assessment tools that were used to **specifically assess laparoscopic suturing in the operating-room**

	OSATS-GRS	GOALS	
	Antosh 2013	Original	Van Sickle LS error rating
CONTENT			
Expert judgment			
Task analysis			
Consensus method			
RESPONSE PROCESS			
Rater training			+
Score interpretation and meaning			
INTERNAL STRUCTURE			
Rater reliability	+	+	+
Item analysis			
Generalizability theory			
Others			
RELATIONS TO OTHER VARIABLES			
Training level or case experience			+
Other performance assessment tool scores			
Time			
Operative data			
Motion analysis			
Simulator scores			
CONSEQUENCES			
Applications to residency program			
Criterion-referenced score (benchmark or pass/fail)			

OSATS-GRS Objective Structured Assessment of Technical Skills-Global rating scale, GOALS Global Operative Assessment of Laparoscopic Skills, LS Laparoscopic Suturing.

Table 2: Validity evidence for the assessment tools that were used to **specifically assess laparoscopic suturing in simulation setting**

	OSATS-GRS				GOALS			
	Original	Bingener 2008	Broe 2006	Kowalewski 2014	Original	Stelzer 2009	LSIKT-GRS	LSIKT-CL
CONTENT								
Expert judgment							+	+
Task analysis								
Consensus method								
RESPONSE PROCESS								
Rater training	+	+				+		
Score interpretation and meaning								
INTERNAL STRUCTURE								
Rater reliability	+	+	+	+	+	+	+	+
Item analysis	+							
Generalizability theory								
Item response theory								
Others[a]							+	+
RELATIONS TO OTHER VARIABLES								
Training level or case experience	+		+		+	+	+	+
Other performance assessment tool scores	+	+						
Time				+	+			
Operative data								
Motion analysis			+		+	+	+	+
Others[b]				+		+		
CONSEQUENCES								
Applications to residency program								
Criterion-referenced score (benchmark or pass/fail)						+		

	ASVC[c]	ASVC[d]	LS-GRS	LS-QRS	LS-AR	IKT-CL	LS-CL	LS-ERS
CONTENT								
Expert judgment						+	+	
Task analysis								
Consensus method						+	+	
RESPONSE PROCESS								
Rater training								
Score interpretation and meaning								
INTERNAL STRUCTURE								
Rater reliability	+	+			+	+	+	

Item analysis							
Generalizability theory							
Item response theory							
Others[a]							
RELATIONS TO OTHER VARIABLES							
Training level or case experience	+	+	+	+	+		+
Other performance assessment tool scores							
Time							
Operative data							
Motion analysis							
Others[b]					+		
CONSEQUENCES							
Applications to residency program							
Criterion-referenced score (benchmark or pass/fail)		+					

	LS Checklist by Moorthy			Van Sickle LS error rating			
	Original	Munz 2007	Tjiam 2013	Original	Takazawa 2015	ASVC[c]	LVG
CONTENT							
Expert judgment						+	
Task analysis							
Consensus method						+	
RESPONSE PROCESS							
Rater training	+			+	+		
Score interpretation and meaning							
INTERNAL STRUCTURE							
Rater reliability	+	+		+	+	+	+
Item analysis							
Generalizability theory			+				
Item response theory							
Others[a]							+
RELATIONS TO OTHER VARIABLES							
Training level or case experience	+	+		+	+	+	
Other performance assessment tool scores	+						
Time							
Operative data							
Motion analysis	+	+					
Others[b]							

CONSEQUENCES	
Applications to residency program	
Criterion-referenced score (benchmark or pass/fail)	+

OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, CL Checklist, LS Laparoscopic Suturing,
GOALS Global Operative Assessment of Laparoscopic Skills, LSIKT-GRS LS and intracorporeal knot tying-GRS, ASVC Assessment for suturing
of vaginal cuff, LS-QRS LS-Quality rating scale, LS-AR LS in augmented reality simulator, IKT-CL Intracorporeal knot tying-CL,
LS-ERS LS-Error rating scale, LVG Laparoscopic Video Grader.

[a]Others: Inter-station reliability, test-retest reliability

[b]Others include intra-operative assessment tool scores and simulator scores

[c,d,e]There are three different tools for ASVC. [c]Tunitsky-Bitton 2016, [d]King 2015, [e]Weizman 2014

Table 3: Validity evidence for the assessment tools that were used to **assess procedures that required laparoscopic suturing in the operating-room** (e.g. laparoscopic nissen fundoplication). The procedures are stated in table 2.

	OSATS-GRS	GOALS							
	Original	Birkmeyer 2013	Bingener 2008	Hiemstra 2011	Original	Lee 2012	GOALS-GH	GRS-OS	ALRYGB
CONTENT									
Expert judgment							+		
Task analysis									
Consensus method									
RESPONSE PROCESS									
Rater training	+								
Score interpretation and meaning									
INTERNAL STRUCTURE									
Rater reliability	+	+			+	+	+		
Item analysis		+	+		+		+		+
Generalizability theory					+				
Others*									
RELATIONS TO OTHER VARIABLES									
Training level or case experience	+			+	+		+	+	+
Other performance assessment tool scores					+		+		
Time									
Operative data		+			+				
Motion analysis									
Simulator scores						+			
CONSEQUENCES									
Applications to residency program									
Criterion-referenced score (benchmark or pass/fail)									

				OCRS				
	LNF-CL	GRITS	TAPP-CL	Original	Ghaderi 2015	OPRs	BOSATS	GERT
CONTENT								
Expert judgment			+				+	+
Task analysis			+				+	
Consensus method			+				+	

RESPONSE PROCESS								
Rater training	+							+
Score interpretation and meaning								
INTERNAL STRUCTURE								
Rater reliability	+		+			+	+	+
Item analysis		+		+	+			
Generalizability theory								
Others*							+	
RELATIONS TO OTHER VARIABLES								
Training level or case experience		+	+				+	+
Other performance assessment tool scores			+				+	+
Time								
Operative data								
Motion analysis								
Simulator scores								
CONSEQUENCES								
Applications to residency program								
Criterion-referenced score (benchmark or pass/fail)								

OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, GOALS Global Operative Assessment of Laparoscopic Skills, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of Laparoscopic Roux-en Y gastric bypass, BOSATS Bariatric Objective Structured Assessment of Technical Skill, LNF-CL Laparoscopic nissen fundoplication-Checklist, GRITS Global Rating Index for Technical Skills, GERT Generic Error Rating Tool, TAPP-CL Transabdominal Preperitoneal procedure-Checklist, OCRS Objective Component Rating Scale, GOALS-GH Groin Hernia, OPRs Operative Performance Rating System

*Others: Inter-station reliability, test-retest reliability

Table 4: Validity evidence for the assessment tools that were used to **assess procedures that required laparoscopic suturing in simulation setting** (e.g. simulated laparoscopic nissen fundoplication). The procedures are stated in table 2.

	OSATS-GRS								
	Original	Crochet 2016	Dath 2004	GRS-OS	ALRY GB	OCRS[c]	LNF-CL	FLP	LIHR-ERS
CONTENT									
Expert judgment						+			
Task analysis							+		
Consensus method						+			
RESPONSE PROCESS									
Rater training									
Score interpretation and meaning									
INTERNAL STRUCTURE									
Rater reliability	+	+	+	+	+	+			
Item analysis	+								
Generalizability theory									
Item response theory									
Others[a]			+			+			
RELATIONS TO OTHER VARIABLES									
Training level or case experience		+		+	+		+	+	+
Other performance assessment tool scores									
Time									
Operative data									
Motion analysis									
Others[b]					+				
CONSEQUENCES									
Applications to residency program									
Criterion-referenced score (benchmark or pass/fail)									

OSATS-GRS Objective Structured Assessment of Technical Skills-Global rating scale, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of LRYGB, OCRS Objective Component Rating Scale, LNF-CL Laparoscopic nissen fundoplication-Checklist, FLP Fundamentals to laparoscopic procedures, LIHR-ERS Laparoscopic inguinal hernia repair-Error rating scale.
For Chang 2015, Barussaud 2015, Zhao 2012, Adrales 2003, and Aggarwal 2007, there are 2 assessment tools that were validated in each of the 5 papers. Therefore, unless specified, the validity will be for both tools.
[a]Others: Inter-station reliability, test-retest reliability
[b]Others include intra-operative assessment tool scores and simulator scores
[c]Original OCRS

Appendix 4.2.5: Guideline for the selection of an assessment tool that was used in the simulation setting

Table 1: Simulation setting: Guideline for the selection of an assessment tool that was used to **specifically assess laparoscopic suturing skills**

			GRS	Checklist	Error rating scale
Recorded	**Reviewer**	**Generic**	**OSATS**(Bingener 2008*, Broe 2006*, Kowalewski 2014*)		
			LVG*		
			GOALS(original*, Stelzer 2009*)		
			LSIKT-GRS**		
		Specific	**LS-AR***		
			LS-GRS*		
			LS-QRS*		
				Moorthy CL (original**, Munz 2007*, Tjiam 2013**)	
				LSIKT-CL**	
				ASVC[a]*	
					Van Sickle LS error rating (original**, Takazawa 2015*)
					LS-ERS*
		Hybrid	**ASVC[b]***		
			ASVC[c]*		
Direct	**Observer**	**Generic**	**OSATS**(original)***		
			LSIKT-GRS*		
		Specific		**LSIKT-CL***	
				IKT-CL*	
				LS-CL*	
				Moorthy CL(original)*	

GRS Global rating scale, CL Checklist, LS Laparoscopic Suturing, OSATS-GRS Objective Structured Assessment of Technical Skills-Global Rating Scale, CL Checklist, LVG Laparoscopic Video Grader, GOALS Global Operative Assessment of Laparoscopic Skills, LSIKT-GRS LS and intracorporeal knot tying-GRS, ASVC Assessment for suturing of vaginal cuff, LS-QRS LS-Quality rating scale, LS-AR LS in augmented reality simulator, IKT-CL Intracorporeal knot tying-CL, LS-ERS LS-Error rating scale.

[a,b,c]There are three different tools for ASVC [a]Weizman 2014, [b]King 2015, [c]T-B Tunitsky-Bitton 2016

*limited (1-5) **moderate (6-10) ***strong level of validity evidence (11-15)

Table 2: Simulation setting: Guideline for the selection of an assessment tool that was used **to assess procedures that required laparoscopic suturing** (e.g. simulated laparoscopic nissen fundoplication)

			GRS	Checklist	Error rating scale
Recorded	**Reviewer**	**Generic**	**OSATS** (Crochet 2016*, Dath 2004*)		
			FLP*		
			GRS-OS*		
		Specific	**ALRYGB***		
			OCRS(original)*		
					LIHR-ERS*
Direct	**Observer**	**Generic**	**OSATS**(original)*		
		Specific		**LNF-CL***	

GRS Global rating scale, CL Checklist, LS Laparoscopic Suturing, OSATS Objective Structured Assessment of Technical Skills, GRS-OS Global rating scale of operative skill, ALRYGB Assessment of LRYGB, OCRS Objective Component Rating Scale, LNF-CL Laparoscopic nissen fundoplication-Checklist, FLP Fundamentals to laparoscopic procedures, LIHR-ERS Laparoscopic inguinal hernia repair-Error rating scale.
*limited (1-5) **moderate (6-10) ***strong level of validity evidence (11-15)