Wrexham University Research Online

Journal Article

Does post-graduate surgical simulation-based education correspond to transfer of skills to real life clinical practice? A systematic review

Kumar, B., Kumar, G., Roberts, R., Hughes, S and Payne

Review articles

Does post-graduate surgical simulation-based education correspond to transfer of skills to real life clinical practice? A systematic review

Bidyut Kumar, MBBS, DGO, MD, FRCOG, MA[1], Geeta Kumar, MD, FRCOG, SFFMLM[1],
Ruth Roberts, MBBS, PhD, MRCOG[1], Stephen Hughes, BSc (Hons), PgCert, MPhil, PhD, CSci, FIBMS FHEA[2],
Joshua Payne, MSc, PhD, MBPSs, FHEA[3]

[1] Obstetrics & Gynaecology, Wrexham Maelor Hospital, [2] Maelor Academic Unit, Wrexham Maelor Hospital, [3] Social & Life sciences, Wrexham University

International Journal of Surgical Education

Background

The effectiveness of simulation training in improving practice in real life has been questioned by many because of lack of good quality evidence.
The purpose of this study was to undertake a systematic review of randomized trials in surgical simulation to find out if simulation-based education really leads to improvement in real life surgical practice.

Methods

Searched published literature between 2000 and 2020. Relevant papers scrutinized to identify work, which fulfilled the criteria for this systematic review. Of 157 abstracts, nineteen papers were selected. Project registration number was REG289 at Edgehill University.

Results

There was heterogeneity in methods of simulation training, outcome measures and assessment technique, making comparison difficult.
Mean error rates, surgical time, objective structured assessment of technical skills (OSATS) scores and Global Operative Assessment of Laparoscopic Skills (GOALS) scores were significantly better in simulation-trained group. Global Rating Scale scores were better in simulation trained group but the improvement was not significant. There was evidence of possible publication bias for some of the outcome measures.

Conclusion

Overall, there was evidence of improved competence in real life practice in the group who underwent targeted simulation-based education in comparison to control groups. Small sample size in majority of trials, variation in technique of simulation training, inconsistencies in assessment and heterogeneity of outcome measures made it difficult to compare results of trials included in this review.

INTRODUCTION AND BACKGROUND

Traditionally training in surgical specialties relied on apprenticeship where experience and expertise was gained by working under a senior experienced surgeon. Such training would often continue for a prolonged period until the trainee felt comfortable and confident in carrying out the procedure on their own. In the absence of a formal structured training programme, graduation from an apprenticeship to a more senior grade would predominantly depend on the opinion of the supervising trainer. In 1889 William Stewart Halstead had formulated the old adage, 'see one, do one, teach one' which was based on the quantum and inten-sity of workload and the trainee doctor personally attending the workplace to care for and treat the real patient.[1] In many countries uninfluenced by the European Working Time Directive, this system of learning remains the mainstay of training, particularly in surgical disciplines where manual dexterity and skills may take a relatively longer time to master.

In the last two- or three-decades laparoscopic surgery has become widely used and has replaced open surgical procedures. This has given rise to novel challenges in education and training of new surgeons. These challenges include the use of long rigid instruments which are much less flexible, have reduced range of movement compared to

hands and fingers, amplify movement and tremor, lack the tactile sensation and also make it difficult to perceive depth and carry out three-dimensional movement.

Restricted working hours as a result of European working time directive, increased demands due to revised surgical waiting list targets, rising cost of operating time on surgical operating lists, heightened public awareness leading to concerns being raised about the ethics of basic surgical training on real patients has led to rethinking of the way surgical training is provided. Bridges and Diamond estimates that financial cost of utilising real life operating room time for training surgical residents in United States is $53 million per year.[2] Scott et al estimate that the cost of training in the Guided Endoscopic Module (Karl Storz Endoscopy, Culver City, CA) varies from $215,000 to $285,000, depending on the quality of video-imaging equipment installed.[3] At the University of Texas South-Western Medical Center, the cost of training residents using a video-trainer is $270 per graduating resident.[3] This value stands in contrast to the cost estimate by Bridges and Diamond at the University of Tennessee Medical Center of using operating room time to train residents for approximately $48,000 per graduating resident.[2] Therefore, training outside of the operating room using simulation-based training (SBT), although expensive, seems cost effective.

However, assessment of cost effectiveness of simulation-based training (SBT) is very difficult and complex to undertake mainly because of the uncertainties associated with assessing cost benefits that arise from learners being able to transfer their skills to real life practice and considering patient outcomes.[4,5]

In the published literature there is no clear and definitive indication of cost analysis of SBT in surgical practice.

The above-mentioned factors have produced changes in surgical training curriculum and methods of assessment of training giving rise to surgical skills training programmes and simulation laboratories. Many researchers believe that simulation-based education can improve surgical skills in both simulation laboratory and in real life surgical operation.[3,6-16] However, many of the assessments following simulation-based training has been undertaken either on cadavers or animal models.[11,17-19] Some authors assessed transfer of training and measured transfer effectiveness ratio in their randomized controlled trial but on scrutiny it became obvious that they had assessed the outcome of their intervention by trainee performance on cadavers.[20]

We excluded internal medicine subjects and diagnostic endoscopies because we wanted to focus on skills transfer in surgical procedures where manual dexterity and expert operative competence comes in to reckoning. Moreover, complexities in surgical care are relatively more, the nature of treatment mostly definitive and irreversible and chances of life changing error higher. Some authors state that amongst reported errors about half the adverse events were surgical in nature.[21,22]

Despite advances in structured surgical training, technical proficiency remains poorly evaluated. In present day surgical training programmes at no point is there a compulsion for objective assessment of technical skills. Evaluation of competence is very prone to subjectivity and bias in assessment. Only very few studies have been carried out to show skills transfer from simulation room to the real-life operating scenario.

AIMS AND OBJECTIVES

This systematic review of the randomised trials aims to find out if such resource worthy simulation-based training leads to any benefits in actual real life surgical practice.

From practical personal experience about assessment of trainees the authors were wary of the variation in nature and type of assessment applied and their subjectivity. We aimed to find any difference in speed or duration of surgical procedure, and any identifiable improvement in assessment of technical skills by utilising any of the commonly used assessment tools, for example, Objective Structured Assessment of Technical Skills (OSATS). Our target group were learners who underwent SBT followed by operating in the real-life scenario. The authors also wanted to identify any possibility of bias in reporting as well as the statistical power of the respective trials.

METHODOLOGY

Because of the rapid development and advancement in technology in simulation-based education in the present millennium, we decided to restrict our review to the published literature between 2000 and 2020. We excluded trials where participants were medical students because we wanted to include participants who were committed to a career in surgery and hence had the motivation and desire to improve and do better in their career.

LITERATURE SEARCH

A search of the published literature was carried out using Medical Subject Headings (MeSH), 'high fidelity simulation training', 'simulation training', 'surgical procedures', 'surgery', 'general surgery', 'clinical competence', 'clinical skill', 'fidelity', 'realism', 'reality', 'accuracy'. The databases searched were, OVID (Embase, American Psychological Association Psycinfo, Medline), Pubmed, Cumulative Index to Nursing and Allied Health Literature (CINAHL), and Cochrane Library. Duplicates were excluded in each database.

The reference list in the latest Cochrane database systematic review (CDSR) 'Laparoscopic surgical box model training for surgical trainees with limited prior laparoscopic experience' was also searched for relevant papers for this systematic review.[23]

The International Standard Randomised Controlled Trial Number (ISRCTN) was searched for any studies registered that conformed to criteria for our systematic review.

A publication in Cochrane Database of Systematic Reviews by Gurusamy KS, Nagendran M, Toon CD, Davidson BR, 2014, Issue 3. Art No.CD010478[23] and the other nineteen papers which were finally selected were searched for cross reference. Reference list in these publications was

Table 1. Criteria for systematic review

Method	Systematic review of published English literature between year 2000 and 2020. PICO (Population, Intervention, Comparison and Outcomes) tool was used.
Inclusion criteria (if applicable)	Any surgical discipline, e.g., General surgery, orthopaedics, Obstetrics and Gynaecology, ENT, Ophthalmology, maxillo-facial and others Post-graduate learners Surgical procedure skills English language publications (to avoid translation bias and expenditure thereof) Published between 2000 and 2020 Ethics approval and/or formal consent process declared Randomized controlled trial Assessment or evaluation of simulation training carried out on real life patients in operating room
Exclusion criteria (if applicable)	Non-clinical subject e.g., Anaesthesiology, Physiotherapy, Dental Non-surgical discipline e.g., Cardiology, Resuscitation Diagnostic endoscopy Non-technical skills Veterinary practice Undergraduate students Any language other than English language publication (in order to avoid translation bias) Non-randomized studies, prospective cohort studies or retrospective studies. Where trial registration or ethics approval was not explicit Assessment or evaluation undertaken on simulators (including cadaver)
Proposed data synthesis/ analysis	Standard systematic approach. Scrutiny of trial method, technique of assessment and outcome measures used. Meta-analysis, by production of forest plots on outcome measures and funnel plots for assessment of publication bias.

Summary of procedure used for database search

OVID	55 papers	**Two papers were selected.** **The remaining 53 papers were excluded because of following reasons. In thirty papers assessment of skills transfer was tested on surgical simulator or on cadaver. Twelve papers were about different subject matter. One paper was a review essay. Four papers were result of survey of participants and in six papers the trials were non-randomized study**
PubMed	6 papers	Two papers which were short listed, were duplicate with papers found in OVID database. Of the remaining, one paper was a narrative review, and in the other three papers authors had assessed participants during simulated procedures
Cochrane Library	58 papers	Of these eight papers were shortlisted. In this selection there was two duplicates with papers obtained from OVID database. So, **six papers were selected** from this database. Those not selected were for the following reason. Seventeen dealt with different subject matter. In twenty-six papers, the assessment of training was carried out on simulators or cadaver or animal models. In one paper, the participants were medical students. Four were protocols where final papers were not published. One paper was on prospective controlled trial. One paper was a conference abstract which was later published by another author in the same group which presented at the conference.
CINAHL	8 papers	Of these three were short listed, two of which were duplicates with papers selected from OVID database and one duplicate with search in Cochrane library
ISRCTN	13 papers	Of these five papers were short listed, two of which were duplicates of papers selected from other databases. Two other authors were contacted via e-mail, but no response was received. One paper was published after October 2021 The reasons for not selecting the remaining eight papers were as follows. Participants were medical students in two papers. Two papers were on different subject matter. In two papers the assessment of learners was done on simulators or cadavers. Two study result had not been published. Other cross reference within this group yielded three papers of which **two papers were selected**. The one that wasn't selected was because participants were not randomized in that study.

searched and abstracts were reviewed in twenty-four selected papers. Of these fourteen papers were short listed. Of these fourteen papers, two abstracts were conference papers which could not be obtained. There was no response from these two authors following our e-mail communication.

So, twelve papers were identified and selected from cross reference of this Cochrane Systematic Review and those finally selected nineteen papers. One of these papers was a review essay, and two papers were about non-randomized trial. **So, nine papers were selected** from these cross references.

A total of 157 abstracts were reviewed by BK, RR and GK

Nineteen papers were finally selected for this systematic review (n=19).

```
┌─────────────────────────┐
│ Following literature     │
│ search and               │
│ duplicates removed       │
│ 157 abstracts reviewed   │
└─────────────────────────┘
```

┌──────────────┐ ┌──────────┐ ┌──────────┐ ┌──────────────┐
│ Papers on   │ │ Review   │ │ Trial on │ │ Assessed on │
│ different   │ │ essay 3 │ │ Medical │ │ simulators/ │
│ subject   │ │   │ │ students │ │ cadaver/other│
│ matters on   │ │   │ │   3   │ │ models 67   │
│ simulation   │ │   │ │   │ │   │
│ 40   │ │   │ │   │ │   │
└──────────────┘ └──────────┘ └──────────┘ └──────────────┘

┌──────────────┐
│ Survey of   │
│ participants 4│
└──────────────┘

┌──────────────────────┐
│ Full papers reviewed 40│
└──────────────────────┘

┌──────────────────────┐   ┌──────────────────────┐
│ Non-randomized trials 9│   │ Results not published 12│
└──────────────────────┘   └──────────────────────┘

┌──────────────────────────┐
│ Papers included in this │
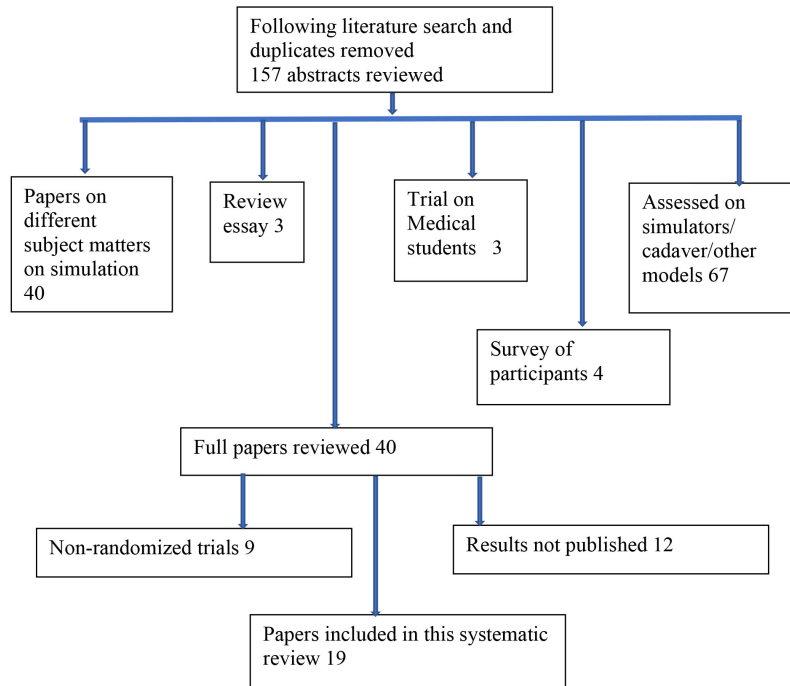│ systematic review 19   │
└──────────────────────────┘

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram showing selection of papers for this systematic review

DATA EXTRACTION AND ANALYSIS

BK extracted all the data from the nineteen included trials and was checked by RR using standard data extraction tables agreed and produced by the authors. Each included trial was scrutinized and critically appraised for its quality and details. The quality assessment of these trials included, method of reporting, methods of randomisation, sample sizes, allocation concealment, blinding of assessors and interrater reliability of assessors.

Pooling of data from these nineteen trials was not possible due to the heterogeneity of methods used and the variety of outcome measures used.

A total of 157 abstracts were reviewed by BK, RR and GK

Nineteen papers were finally selected for this systematic review (n=19).

RESULTS

ETHICS APPROVAL AND CONSENT

Fifteen out of the nineteen trials declared obtaining ethics approval from their respective organisational department and consent from the participants. In their published paper, Banks et al,[24] Larsen et al,[25] and Seymour et al,[6] did not explicitly state about ethics approval but their trials were registered and as such would have included ethics approval. Zandejas et al[26] stated that their trial was found to be exempt from requirement of ethics approval by the Mayo Clinic Institutional Review Board, Rochester, MN, USA. All their trial participants gave consent to be part of the trial.

RANDOMIZATION, BLINDING AND TYPE OF ASSESSMENT

Authors of all the included trials declared having undertaken random allocation of participants in to SBT and Control groups. Fourteen of nineteen authors also stated the exact method of randomisation. Some of the authors, while making it clear that they undertook randomization, didn't actually mention the exact methods that they followed.[6, 27-30]

With the exception of one trial by Gala et al, rest of the assessors were blinded to the randomization status of the participants.[31] In another trial by Hogle et al, assessors present in operating room (OR) were unblinded but in the same trial video assessors were blinded to the training status of participants.[32] In ten of the nineteen trials assessors were present in OR and in the remaining trials assessment was performed on video recordings of the operative procedures carried out by participants. In two of the assessments in OR there was additional assessment of video recording of the same procedure by other assessors.[28,32] This data is presented in Tables 3-10.

SURGICAL PROCEDURES TESTED, TYPE OF SIMULATOR USED AND COMPARATOR

The terminologies, 'trainees' or 'participants' has been used to refer to individuals who were the test subjects in the trials included in this systematic review. The group, which received focussed or targeted training have been referred to as the Simulation-based training (SBT) group. The comparator group has been referred to as the Control group who either received no SBT or had access to SBT but with-

out any formalized targeted plan or predefined skills targets to be achieved.

Seven RCT s tested skills transfer in laparoscopic cholecystectomy procedures, four RCT s tested skills transfer in laparoscopic salpingectomy, three in laparoscopic total extraperitoneal hernia repair, two each in laparoscopic tubal ligation and knee arthroscopy and one for intracorporeal suturing and knot tying in a laparoscopic Nissen fundoplication procedure. All these RCT s compared surgical performance of participants who underwent SBT with the control group. There was a wide variation of simulators used to train the participants in the different trials.

LapSim Virtual Reality Simulator was used in four trials.[25,28,32,33] Minimally Invasive Surgical Trainer Virtual Reality (MIST VR) was used in three trials,[6,9,30] and standard laparoscopic simulator was used in two trials[24,31]

Video laparoscopy training was used in two trials.[27,34] The remaining trials used one of the following simulators: ArthroSim Virtual Reality Simulator, moulded rubber hernia simulator, McGill Laparoscopic Inguinal Hernia Simulator, Video and porcine cadaver, dry knee arthroscopy model, Box trainer and Virtual Reality Simulator, Fundamentals of laparoscopic surgery programme and Web-based Mastery Learning. This is shown in Table 2.

As shown in Table 2, in vast majority of trials, participants in the comparator group or controls were not allowed access to SBT. Only in four of the nineteen trials the participants randomized to Control group were allowed to practice on simulators although their practice was not formalised, was not supervised nor did they have any predefined targets to achieve before their skills assessment[27,28,30,35]

BASELINE ASSESSMENT OF PARTICIPANTS OF SIMULATION-BASED TRAINING GROUP AND CONTROL GROUP

In all the nineteen trials, a baseline assessment of the SBT group and Control group was carried out. Kurashima et al. declared that their control group participants had slightly more overall laparoscopic experience than SBT group at baseline assessment (p=0.045).[36] Remaining characteristics of the two groups were not significantly different (p>0.05). In the remaining eighteen trials the demographic characteristics, relevant operative skill and past experience in the relevant surgical procedure were similar, and not statistically different amongst the SBT and Control groups (p >0.05). In the trials where a mixture of junior and senior residents was recruited, the authors carried out block randomization so that the composition and mixture of each group, SBT and Controls, were similar.[36,37] Sroka et al state that they carried out a baseline assessment of their recruited participants using GOALS score and excluded from randomization all those participants who scored more than 15.[38]

DURATION OF STUDY AND TIME TO ASSESSMENT

The elapsed time after training was completed and before the assessment was undertaken was not very clear in all the trials. In almost trials the final assessment seems to

have taken place within a short time after the training of the SBT group participants was completed. The maximum duration to assessment appears to be 316 days but this was almost similar in both the SBT group and the Control group in this particular trial.[31] In all trials, the assessment on real patients was undertaken after the predefined proficiency was attained by the participants in the SBT group. In most trials, the duration of study and time to assessment was dictated by the residency rotation programme ^24, 30, 34, 36,38, 39,^ . Table 2 demonstrates the time utilised and duration to assessment of competence.

EXPERIENCE AND TYPE OF PARTICIPANTS

In fifteen of the nineteen trials in this systematic review, the participants were postgraduate residents ranging from year 1 to year 4 without any previous experience of the operative procedure that was being assessed. Kurashima et al, Van Sickle et al and Zandejas had recruited some postgraduate years 5 or 6 but these residents did not have prior experience of the operation being tested.[26,30,36] Grantcharov recruited sixteen surgeons for their trial on laparoscopic cholecystectomy, but these surgeons had 'limited' experience in this procedure and their SBT group matched with their controls.[9]

Coleman et al found that in their trial the improvement of performance in laparoscopic partial salpingectomy was most striking in postgraduate year 3 residents compared to postgraduate year 4 residents.[27]

TYPE OF OPERATIVE CASES AND ROLE OF SUPERVISOR AND ASSESSOR

In all trials, simple and uncomplicated patients were chosen for operative assessment. When any complication was detected on the operating table, those cases or participants were excluded from trial. In vast majority of the trials, authors have not mentioned about any instance of take-over by the supervisor or experienced assistant. Ahlberg et al mentions that in their trial, three operations had to be converted from planned laparoscopic cholecystectomy to open surgical procedure.[33] Cannon et al mentions that in their trial if the resident participant took more than 25 minutes then the attending supervisor took over the operative procedure and the unfinished task in that instance was allocated a score of zero.[39] In many of the trials, an experienced surgeon supervised and assisted in the surgical procedure but the assessment and scoring for the trial was done by an independent blinded assessor.[9,25,28,29,32,34,36,40]

OUTCOME MEASURES USED AND FINDINGS

Different outcome measures were used by the trial authors for assessment of transfer of skills from SBT to real life practice, as shown in Tables 3-10. In these tables, the group of participants who had simulation-based training (SBT) has been referred to as SBT group and the comparator group who did not have any targeted simulation training has been referred to as the Control group of C group.

Table 2. Type of training, comparator, time to assessment, power calculations

Author and year of publication. Power calculation	Grade and number of subjects and Surgical procedure tested	Training method (number of participants)	Comparator / Control (Number of participants)	Time to assessment
Ahlberg, 2007[33]	PGY 1 to 2. 13 residents. Laparoscopic cholecystectomy. Assessors blinded	LapSim VR Simulator (7)	Standard residency training. No SBT (6)	First procedure within 2 weeks for controls. Training group first procedure after attaining proficiency. All procedures within 6 months.
Banks, 2007[24]	PGY 1. 20 residents. Laparoscopic bilateral tubal ligation	Standard Lap Simulator made by Limbs and Things, UK (10)	OR training alone. No SBT (10)	Training and assessment within 4 months of study
Cannon, 2014[39]	PGY 3. 48 residents. Diagnostic knee arthroscopy.	ArthroSim VR arthroscopic knee simulator (27)	Institutional based traditional training. No SBT (21)	Within 14 days after attaining proficiency
Coleman, 2002[27]	PGY 3-4. 18 residents. Laparoscopic partial salpingectomy.	Video laparoscopic training modules (11)	Allowed to practice in skills laboratory (7)	Week 4 after commencement of study
Gala, 2013[31] For 2-sided alpha of 0.05 and beta of 0.20, a total of 110 participants were needed to show a 50% in improvement.	PGY 1-4. 102 residents. Bilateral tubal ligation.	Faculty directed SBT (48) on laparoscopic simulator	Traditional teaching. NO SBT (54)	Maximum duration to assessment was 316 days
Gauger, 2010[28]	PGY 1. 14 residents. Lap cholecystectomy	LapSim with specific proficiency targets (7)	LapSim but without any specific targets (7)	2 months to assessment after 4 months of training.
Grantcharov, 2004[9]	16 surgeons with limited experience in laparoscopic surgery. Lap Cholecystectomy.	VR training on Minimally Invasive Surgical Trainer Virtual Reality (MIST-VR) (8)	No SBT (8)	Assessment within 14 days after initial assessment
Hamilton, 2001[29]	PGY 3-4. 21 residents. Laparoscopic TEP hernia repair	Training on molded rubber hernia simulator and interactive CD-ROM (10)	Traditional OR training. No SBT (11)	Assessment during and within 2 weeks study period
Hogle, 2009[32]	PGY 1. 12 residents. Lap cholecystectomy	LapSim VR Simulator (6)	Traditional OR training. No SBT (6)	Assessment at the end of 5-week period
Kurashima, 2014[36] To show significant difference between the two groups with an alpha of 0.05 and beta error of 0.20, at least 7 participants in each group needed. Basis of power calculation was a mean GOALS score of 20+/- 3 in the OR but achieved 18.2.	PGY 2-5. 14 residents. Laparoscopic TEP hernia repair	McGill Laparoscopic Inguinal Hernia simulator (5)	Standard traditional residency (9)	Residents tested after 3 months of baseline assessment. Training group had an additional assessment within 2 weeks of achieving proficiency.

Author and year of publication. Power calculation	Grade and number of subjects and Surgical procedure tested	Training method (number of participants)	Comparator / Control (Number of participants)	Time to assessment
Larsen, 2009[25] With an alpha of 0.05 (2-sided) and power 80% (largest SD 4.40) 18 or more participants required.	PGY 1-2 21 residents. Laparoscopic salpingectomy	LapSim (11)	No SBT (10)	Assessment soon after SBE was complete
Patel, 2016[37] 80% power to identify improvement of 0.75 in mean score with a SD of 25% of the mean score in both study groups. Sample=11 in each group	PGY 1 22 residents. Laparoscopic salpingectomy	Simulation on video and porcine cadaver (11)	Standard OR training (11)	Assessment within 90 days
Roberts, 2019[40] 15 participants per group required for a 90% chance (beta=1) with alpha of 0.05 (2 tailed test).	PGY 2-3 30 residents. Diagnostic knee arthroscopy.	Simulation on dry knee arthroscopy model and American Board approved simulator (15)	Standard residency training. No SBT (13)	Study period 13 weeks.
Scott, 2000[34] Power =0.8 with a type 1 error of 0.05, sample size calculated was 27. But 5 drop-outs meant sample of 22.	PGY 2-3 22 residents. Laparoscopic cholecystectomy	Video trainer (9)	No SBT (13)	Assessment at the end of 1 month of residency
Seymour, 2002[6]	PGY 1-4 16 residents. Laparoscopic cholecystectomy	VR training. MIST VR. (8)	Standard residency. No SBT (8)	Training session lasted approximately 1 hour
Shore, 2016[35] For an alpha score of 0.05 and a power of 0.8, sample size for a 2-sided test was 10 in each treatment arm.	PGY 1-2 27 residents. Laparoscopic salpingectomy	Box trainer and VR simulator (14)	Conventional training which could include simulation. Simulation was not targeted (13)	Tested within the 7-week study period
Van Sickle, 2008[30]	PGY 3-6. 22 residents. Nissen fundoplication-fundal intracorporeal suturing part of the procedure	MIST VR plus box trainer, supervised training (11)	Standard residency and access to SBT without supervision (11)	Rotation lasted 4 to 5 weeks for PGY 3 and 8 weeks for PGY-5 residents.
Sroka, 2010[38] Calculated those 7 subjects in each group= a power of 80% to detect a difference of 5 points in GOALS score	PGY 16 residents. Laparoscopic cholecystectomy.	Fundamentals of laparoscopic surgery programme, American College of Surgeons (8)	Standard residency. No SBT. (8)	After at least 6 weeks for Control group and for SBT group after proficiency was attained
Zandejas, 2011[26] 25 residents per arm= 80% power to detect a 5-minute decrease in operative time with an alpha level of 0.05	PGY 1-5 50 residents. TEP hernia repair	Web based mastery learning curriculum followed by TEP Simulator (26)	Standard residency programme. No SBT (24)	Study period 20 days between baseline assessment and post-training assessment.

Abbreviations: PGY: Postgraduate year; SBT: Simulation based Training; TEP: Total Extraperitoneal; MIST VR: Minimally Invasive Surgical Trainer Virtual Reality; VR: Virtual reality

Ahlberg,[33] Seymour[6] and Van Sickle[30] evaluated the participant's performance using mean number of errors, surgical time, and other measures like conversion of laparoscopic procedure to open surgery and excess needle manipulation. The group of trainees who had SBE made significantly smaller number of errors (p= <0.05). Surgical time was also significantly less in the SBT group (p=<0.05). Conversion of laparoscopic procedure to open surgical procedure was required thrice in the Control group compared to none in the SBE group.[33] Van Sickle found that the SBE group in their trial carried out significantly less needle manipulation than their controls (p= <0.05).[30] Details are presented in Table 3.

As shown in Table 4, some authors used one or another type of Global Rating Scale (GRS) on its own to assess transfer of skills to operation on real life patient.[27,32,34,36,38] These GRS were devised by different authors as indicated in Table 4. The SBE group performed significantly better (p= <0.05) than the control group in all but one of these trials. Hogle found that in their trial, in the domains of depth perception, bimanual dexterity, efficiency, tissue handling and autonomy, there was no significant difference between the performance of SBT group and Control group, although the SBT group fared better than the Control group (p= 0.55 to 0.99).[32]

Banks, as shown in Table 5, used a 25-point Task specific check list, a GRS and pass rate to assess the participants in their trial. In all these three outcomes measures their SBE group performed significantly better than their Control group (p= 0.002, 0.003 and 0.003 respectively)[24]

Cannon et al, in their trial of diagnostic knee arthroscopy used a procedural checklist, visualization scale, probing scale and a global rating to assess performance of participants in their trial. In all these outcome measures, except the visualisation scale, their SBT group were found to be significantly better that their control group (p= 0.031, p=0.34, p= 0.016, p=0.061 respectively). In the visualisation scale the SBE group performed better but not significantly so.[39] Table 6 shows the findings of this trial.

Table 7 demonstrates the trial results of those authors who used Objective Structured Assessment of Technical Skills (OSATS) as the outcome measure to compare SBE group with their control group.[25,31,35,37] The OSATS tools used in these trials were devised by different authors as shown in table 8. In all these trails the SBE group performed significantly better than the control group (p=< 0.05)

Gauger,[28] Hamilton[29] and Zandejas,[26] used Global Operative Assessment of Laparoscopic Skills (GOALS) to assess and evaluate the performance of participants in their trials. The authors of GOALS tool, which these authors used have been mentioned in Table 8. The SBT group in all three trials performed much better than the Control group.

However, the improvement seen by Gauger et al was not statistically significant (Overall competence p= 0.228 and task completion score, p=0.345).[28] They found significantly lower number of errors in their SBT group. In the trials by Hamilton[29] and Zandejas,[26] GAOLS score was significantly better in the SBT group (p= < 0.05). In addition, instrument knowledge and handling were significantly better in Hamilton's trial.

The trial by Zandejas,[26] was the only one amongst the nineteen trials included in this systematic review, where patient related outcomes were investigated. Apart from significantly better operative time (p=0.0001), Zandejas found that intraoperative complications, postoperative complications and overnight stay were significantly less likely in the SBT group (p= < 0.05).

Grantcharov found that their SBE group performed significantly better than their control group in the outcomes they studied, that is, economy of movement (p=0.003), duration of procedure (p=0.021) and error score (p= 0.003).[9] Table 10 shows the findings.

The trial by Roberts et al was novel in that they used wireless elbow worn motion sensors to surgical performance objectively.[40] Their primary outcome measure for diagnostic knee arthroscopy was number of hand movements where the SBT group performed significantly better than the control group (p= < 0.001). For their secondary outcomes (minor movements, smoothness and time taken) the SBT group was significantly better than the control group (p= < 0.001). Average time taken by SBE group was 320 seconds versus 573 seconds by the controls (p= < 0.001).

COMFORT LEVEL OF PARTICIPANTS IN THE TRIALS

In four of the nineteen trials authors looked at comfort levels or perception of participants involved in the trial. Coleman (2002) found that both their SBE and Control groups expressed lower than average comfort levels pre-test and post-test.[27] The pretest scores for SBT group were 12 versus 13.5 for controls and post-test scores were 16 versus 17 respectively. There was statistically significant improvement in the post test comfort levels (p=0.001). There was no difference between the SBE and control groups. Within the SBE group there was a weak association between self-perception and operative evaluation by Global Surgical Assessment Tool (GSAT).

Patel (2016) found that in their control group there was no change in 8 of the 10 subjective comfort levels.[37] In their SBT group there was an improved sense of comfort with anatomy knowledge (p=0.02), surgical steps (p=0.004), double handed surgical technique (p=0.04), knowledge about energy (p=0.002), understanding of the risk-benefit of the procedure (p=0.04). They also perceived benefit of viewing procedural video and SBT before performing the real operation.

In their trial, Scott (2000) found that 3 of 13 of their control group participants and 5 of 9 SBT group participants felt comfortable with their laparoscopy skills.[34] After completion of the study, their perception had improved to 6 of 13 and 8 of 9 in the respective groups. Amongst those who didn't feel comfortable 3 of 10 controls and 3 of 4 SBE group participants perceived a sense of comfort after the study (p= 0.175). In the SBE group all nine participants felt that the video trainer was useful and eight of the nine felt that the training had enhanced their skills during real life operating.

Table 3. Trials assessing error rates, surgical time and others

Author and year. Mode of assessment. Blinding status	Result Power calculations	Mean and variance of errors	Surgical time	Conversion to open surgery/ Other measures
Ahlberg, 2007[33] Video Blinded	SBT significantly better than C Based on Means and SD s, the statistical power was 0.999	Errors for entire procedure, C 86.2 vs SBT 28.4 (95 CI, C 58.18 to 114.12, V=916.68; SBT 23.51 to 33.32, V 118.69) (p=0.0037) Exposure errors, C 53.4 vs SBT 15.4 (95 CI, C 16.70 to 90.13, V 623.31; SBT 11.16 to 18.79, V 68.44) (p=0.0402) Clipping & tissue division, C 7.1 vs SBT 1.9 (95 CI, C 3.95 to 10.25, V 41.11; SBT 0.93 to 2.87, V 5.57) (p=0.0080) Dissection errors, C 29.5 vs SBT 11.5 (95 CI, C 13.99 to 45.01, V 61.50; SBT 8.82 to 14.08, V 28.77) (p=0.0310)	58% longer in C group. Difference not significant (p=<0.0586)	Conversion: 3 in C group. 0 in T group
Van Sickle, 2008[30] video Blinded	Intracorporeal suturing and knot tying skills SBT significantly better than C	Suturing errors SBT 25.6+/-9.3 vs C 37.1+/-10.2 (P=<0.01);	Suturing time (sec), SBT 525+/-190 vs C 790+/-171 (p=0.003);	excess needle manipulation, SBT 18.5+/-10.5 vs C 27.3+/-8.6 (p=0.05).
Seymour, 2002[6] Assessed on video recording Blinded	Improvement of all outcome measures in SBT group compared to Control group	Controls made 6 times as many errors as the SBT group with 4 times the variability of SBT as indicated by standard error. Mean number of scored errors per procedure, SBT 1.19 vs Controls 7.38; P=<0.006.	Duration of dissection in SBT was 29% less than in the Controls. NOT statistically significant.	Lack of progress per case, SBT 0.25 vs C 2.19 (p=0.008)

Abbreviations. C: Controls; T: Training group; SBT: Simulation based training; CI: confidence intervals; SD: standard deviation; V: Variance

Van Sickle (2008) reported that several of their participants admitted to feeling nervous in the high-stake environment instituted for the assessment in the trial.[30]

META-ANALYSES

Five meta-analyses were conducted for mean error rates (k = 5), surgical time (k = 7), OSATS (k = 4), GRS (k = 4), and GOALS (k = 4). All meta-analyses were conducted with the MAJOR package (v 1.2.1; Hamilton, 2021) in Jamovi (v 2.2.1; The Jamovi Project, 2021).[41] Standardised mean differences were extracted for all outcome variables given the between-study heterogeneity in procedures and operationalisation of outcomes**,** Standardised mean differences were extracted for all outcome variables given the between-study heterogeneity in procedures and operationalisation of outcomes. Summary effect sizes have been produced from random effects meta-analyses, fit using restricted-maximum likelihood, with the Knapp and Hartung (2003) adjustment applied.[42] The results of these meta-analyses should be interpreted with caution with respect to the discussion around heterogeneity of methods, generally small sample sizes, and potential for publication bias. Model summaries, forest plots, and funnel plots are presented in Figure X. Pro Prospective power analyses were carried out to provide estimates of minimum required sample sizes with the jpower module (version 0.1.2, Morey & Selker, 2019)[43] for Jamovi. A summary of prospective power analyses for all five outcomes is presented in Table 11.

In several papers, complete summary statistics (i.e., means, standard deviations) were not presented. WebPlotDigitizer 4.5 (Rohagti, 2021)[44] was used to extract means and standard deviations from relevant figures (Figure 3, Figure 5, Seymour et al., 2002).[6] Where medians were reported with inter-quartile range (Grantcharov, 2004, Figure 2 & WebPlotDigitizer; Larsen et al., 2009; Shore et al., 2016) ,[9,25,35] means and standard deviations were estimated using the Box-Cox method described by McGrath et al. (2020; https://smcgrath.shinyapps.io/estmeansd/).[45] Kurashima et al. (2014)[36] reported median and range values and the quantile estimation method was used to estimate mean and standard deviation (McGrath et al., 2020).[45] Gauger et al. (2010)[28] did not report standard deviations for either group so pooled standard deviation was substituted based on Cohen's d values and means reported in the paper. Coleman and Muller (2002) did not report any form of variance information in their paper and could not be included in the meta-analysis for Global Rating Scales.[27]

RANDOM EFFECTS MODELS

ERROR RATES

Observed standardized mean differences for error rates ranged from -2.45 to -1.09, with all estimates favouring fewer errors on average in the simulation groups versus controls. The estimated average standardized mean difference based on the random-effects model was $\hat{\mu}$ = -1.38 (95%

Table 4. Outcome measure: Global Rating Scale

Author and year of publication. Mode of assessment. Blinding.	Result	Global rating scale (5-point Likert scale that assesses several aspects of surgical skills)	Other outcome scores/ rating Author of Global Rating Scale
Scott, 2000[34] Assessed in operating room Blinded	Significant improvement in SBT group	Overall performance on Global Assessment median scores 0.2 (25th to 75th % -0.5 to -0.6) in C vs 0.7 (25th to 75% 0.6-1.0) in SBT; P=0.007.	Global rating scale by Reznick and colleagues, 1997 (36)
Coleman, 2002[27] Video Assessors blinded	Significant improvement in SBE group assessed with GSAT	GSAT score: SBT pre-test 17.4 vs post-test 21.7. Mean improvement 4.3 (p= 0.0151) Control pre-test 16.4 vs post-test 20.3 Mean improvement 3.9 (p= 0.0923)	GSAT by Reznick 1997 (36)
Hogle, 2009[32] Assessor in OR unblinded Video assessment blinded	No significant difference between SBT and Control groups	Bimanual dexterity C 2.90+/-0.51 vs SBT 3.17+/-0.42 (p= 0.55) Tissue handling C 3.10+/-0.53 vs SBT 2.96+/-0.59 (p=0.56) Autonomy C 3.11+/-0.62 vs SBT 3.23+/-0.44 (p=0.85) Efficiency C 2.82+/-0.62 vs SBT 2.89+/-0.53 (p=0.93) Depth perception C 3.35+/-0.62 vs SBT 3.60+/-0.55 (p= 0.99)	
Kurashima, 2014[36] Assessors in OR Blinded	Final GOALS scores were higher in SBT group than in the Control group but did not achieve statistical significance	Final GOALS-GH score higher in SBT group (18.2, range 14.9 – 21.5) than C group (14.8, range 12.4-17.1); p=0.06. Total GOALS-GH scores changes, C 1.2 (-1.1 to 3.6) vs SBT 3.4 (2.0 to 4.8) (p=0.08)	GOALS by Kurashima, 2011 (37)
Sroka, 2010[38] Assessed in OR Blinded	Significant improvement in SBT group compared to Controls	Bimanual dexterity C 0.5+/- 1.1 vs SBT 1.25+/- 0.6 (p=0.04) Tissue handling C 0.3+/-0.7 vs SBT 1.13+/-1.0 (p=0.04) Autonomy 0.3+/-1.0 vs SBT 0.6+/-1.1 (p=0.58) Efficiency C 0.4+/-1.1 vs SBT 1.13+/-1.0 (p=0.24) Depth perception C 0.5+/-0.8 vs SBT 1.25+/-0.7 (p=0.08) Total GOALS score of 1.8+/-2.1 in Controls vs 6.1+/-1.3 in SBT; (p=0.0003).	No difference in assessment of difficulty in dissection between SBT and Control groups. Final evaluation: C 2.25 vs SBT 4.5 (p=0.15). GOALS by Vassiliou,2005 (38)

Abbreviation: GOALS: Global Operative Assessment of Laparoscopic Skills; GOALS-GH: Global Operative Assessment of Laparoscopic Skills – Groin Hernia; GSAT: Global Surgical Assessment Tool; SBT: Simulation Based Training

Table 5. Outcome measures: Task specific checklist and Global rating Scale

Author and year of publication. Mode of assessment Blinding	Result	Task specific check list (25-point check list: preoperative, surgical technique, laparoscopic technique, BTL specific skills)	Global rating scale (GRS-5 point Likert scale that assesses 7 aspects of surgical skills) GRS by Reznick, 1997 (36)	Pass rate
Banks, 2007[24] Assessor present in OR Blinded	SBT group better than Control	SBT 92% (SD=7) vs Controls 57% (SD=20) (P= 0.002)	SBT 64% (SD=5) vs Controls 45% (SD=11) (P=0.003)	SBE 100% vs Controls 30% (P=0.003)

Abbreviations: BTL: Bilateral Tubal Ligation; SBT: Simulation Based Training

Table 6. Outcome: Procedural checklist, Visualisation, Probing, GRS

Author and year of publication. Mode of assessment Blinding	Result	Procedural checklist score (% of total possible points)	Visualization scale score (% of total possible points)	Probing scale score (% of total possible points)	Global rating score (% of total possible points)
Cannon, 2014[39] Video Blinded	Training group better than control, but did not reach significance in all scores	SBT mean 63+/-20 vs C 52+/-21 (p=0.031)	SBT mean 61+/-20 vs C 58+/-21 (p=0.34)	SBT mean 64+/-24 vs C 48+/-24 (p=0.016)	SBT 64+/-20 vs C 57+/-19 (p=0.061)

Abbreviation: GRS: Global Rating Scale; SBT: Simulation based training

Table 7. Outcome: Objective Structured Assessment of Technical Skills

Author and year of publication. Mode of assessment Blinding	Result	Before and after Objective Structured Assessment of Technical Skills (OSATS)	After OSATS Control vs SBE group (OSATS author)	Operating time/ other outcome measures
Gala, 2013[31] Assessors present in OR Not blinded	Significant improvement in Simulation-based training (SBT) group	SBT before OSATS 25.6+/-5 vs after OSATS 30.0+/-3 (p= <0.01) C before OSATS 24.7+/-6 vs after OSATS 27.5+/-5, (p= < 0.01)	P= 0.03 (OSATS by Moorthy,2003 (39); Martin et al 1997) (40)	
Larsen, 2009[25] Assessed on video Blinded	Clinically important improvement of operative skills during actual procedure.	Surgical performance total score: SBT Median value 33, range 25-39 IQR 32-36 vs C 23, range 21-28, IQR 22-27. (p=<0.001)	(OSATS by Larsen, 2008) (41)	SBT Median value of 12 mins (range 6-12, IQR 10-14) vs C 24 mins (range 14-38, IQR 20-29) p= <0.001
Patel, 2016[37] Assessed on video Blinded	SBE can improve surgical technique OSAT s	Overall OSAT score SBT pre-intervention of 26.73+/-10.64 vs post-intervention 29.91+/-9.84; (p=<0.001). C pre-intervention 26.64 +/- 10.79 vs postintervention 26.18+/-10.09 (p=0.65)	(OSATS by Niitsu, 2013) (42)	SBE group experienced improved sense of comfort level in most of the 10 subjective comfort levels' P= <0.05.
Shore, 2016[35] Assessment of video recording Blinded	Significant improvement in OR performance of laparoscopic salpingectomy but not for intracorporeal knot tying	SBT, score of 34 (32.25 – 39.25) vs Control score of 30 (27-35) out of maximum of 50, (p=0.043)	(OSATS by Larsen, 2008) (41)	Knot completion SBT 61.5% vs C 45.5% (p=0.431). Knot tying GRS SBT 12.5 points vs C 12 points (p=0.833) Time required to complete knot, SBT 427.5 sec vs C 450 secs (p=0.724)

Abbreviation: IQR-Interquartile range

CI: -1.95 to -0.81) and differed significantly from zero (Figure XA), with evidence of little heterogeneity. Visual inspection of the funnel plot may suggest publication bias, supported by a significant Egger's regression test.

SURGICAL TIME

Observed standardized mean differences for surgical time ranged from -1.41 to -0.25, with all estimates favouring shorter surgical times on average in the simulation groups versus controls. The estimated average standardized mean difference based on the random-effects model was $\widehat{\mu}$ = 1.01 (95% CI: -1.43 to -0.58) and differed significantly from zero (Figure XB), with evidence of little heterogeneity. Visual inspection of the funnel plot may suggest publication bias, but Egger's regression was non-significant.

Table 8. Outcome: Global Operative Assessment of Laparoscopic Skills

Author and year of publication. Mode of assessment Blinding	Result	Global Operative Assessment of Laparoscopic Skills (GOALS)	Average of total number of errors recorded by each reviewer/Other measures/ Operative time	GOALS author
Gauger, 2010[28] Assessors present in OR plus video assessment Blinded	Simulation-based training (SBT) group better than Control, but not statistically significant	Bimanual dexterity C 4.0 vs SBT 4.71 (p= 0.251) Tissue handling C 4.14 vs SBT 5.29 (p=0.091) Autonomy C 3.86 vs SBT 5.0 (p=0.147) Efficiency C 3.71 vs SBT 4.71 (p=0.251) Depth perception C 4.0 vs SBT 5.14 (p= 0.147) Overall competence C 3.71 vs SBT 4.57 (p=0.228)	Total completion score, C 5.43 vs SBT 6.43 (p=0.345) Significantly lower number of erros in the SBT group. C mean 13.64+/- 6.07 vs SBT 6.20+/-3.94 (p=0.004)	Vassiliou 2005 (38)
Hamilton, 2001[29] Assessors in OR Blinded	SBT group significantly better than Control group	Composite post-training score, C 41.0 +/- 23.5 vs SBT, post-training score 65.7+/-17.5 (p=0.01)	Post-training Overall performance GOALS score, C 2.4+/-0.9 vs SBT 3.6+/-0.7 (p=< 0.05)	GOALS by Reznick, 1997 (36)
Zandejas, 2011[26] Assessed in OR Blinded	Significant improvement in SBT (Mastery learning) group compared to Control (Standard learning) group	GOALS scores mean difference between SBT and C = +3.6 (95 CI 2.1 to 5.1) p=0.001.	Operative time: SBT group were on average 6.5 minutes faster, 95 CI = -10.1 to -2.9 (p=0.0001).	Intraoperative, postoperative complications and overnight stay were less likely in the SBT group, OR 0.1, 0.06 and 0 respectively (p= < 0.05) Vassiliou, 2005 (38)

Abbreviation. C-Controls; CI- Confidence intervals

Table 9. Outcome measure: Economy of movement, Duration and Error score

Author, year of publication; Mode of assessment. Blinding	Blinding status	Result	Economy of movement score	Duration of procedure	Error score
Grantcharov 2004[9] Video	Reviewers blinded to training status	Trained group better than Control group	IQR for SBT 2.5 to 5.0 vs C 6 to 7.2 (p= 0.003)	IQR for SBT 45 to 60 mins vs C 45 to 70 mins (p= 0.021)	IQR for SBT 3 to 5.5 vs C 4.5 to 7.2 (p=0.003)

Abbreviation. IQR-Interquartile range

OBJECTIVE STRUCTURED ASSESSMENT OF TECHNICAL SKILLS

Observed standardized mean differences for OSATS ranged from 0.36 to 1.31, with all studies favouring better OSATS outcomes in the simulation trained versus controls. The estimated average standardized mean difference based on the random-effects model was $\hat{\mu}$ = 0.85 (95% CI: 0.34 to 1.37) and differed significantly from zero (Figure XC), with evidence of moderate heterogeneity of study effect sizes. Visual inspection of the funnel plot may suggest publication bias, but Egger's regression was non-significant.

GLOBAL RATING SCALES

Observed standardized mean differences for GRS ranged from 0.35 to 2.33, with all studies favouring better OSATS outcomes in the simulation trained versus controls. The estimated average standardized mean difference based on the random-effects model was $\hat{\mu}$ = 1.35 (95% CI: -0.15 to 2.85)

Table 10. Outcome: Objective Surgical Performance metrics

Author, year of publication. Mode of assessment. Blinding	Result	Objective Surgical Performance metrics (wireless elbow worn motion sensors)	Objective Surgical Performance metrics (wireless elbow worn motion sensors)	Time taken
Roberts, 2019[40] Assessed in OR Blinded	The SBT group consistently outperformed the control group	Primary outcome (number of hand movements): median of 544 (IQR 465 to 593) in SBT vs median of 893 (IQR 747 to 1242) in Controls; p=<0.001.	Secondary outcome: minor movements SBT 176 (IQR 133 to 209) vs C 435 (IQR 310 to 652) p=<0.001. smoothness: SBT 25842 ms^{-3} (IQR 20867 to 27468 ms^{-3} vs C 36846 ms^{-3} (IQR 29840 to 53949 ms^{-3} (p=<0.001).	Time taken SBT 320 secs (IQR 294 to 392 secs vs C 573 secs 9IQR 477 to 860 secs; p=<0.001).

ms^{-3} - meters per second cubed; IQR-Interquartile range

and did not differ significantly from zero (Figure XD), with evidence of moderate heterogeneity of study effect sizes. Visual inspection of the funnel plot suggests publication bias, which was supported by a significance Egger's test.

GOALS

Observed standardized mean differences for GOALS ranged from 0.35 to 2.33, 1.07 to 1.42, with all studies reporting higher GOALS total scores in the simulation trained group compared to controls. The estimated average standardized mean difference based on the random-effects model was $\hat{\mu}$ = 1.20 (95% CI: 0.92 to 1.48) and differed significantly from zero (Figure XE), with evidence of little heterogeneity. Visual inspection of the funnel plot may suggest publication bias, but Egger's regression was non-significant.

PROSPECTIVE POWER ANALYSES

A full review of approaches to power analysis is beyond the scope of this paper but see the recent preprint from Lakens (2022) for pragmatic guidance.[46] Using a data-driven approach, prospective power analyses were conducted for each outcome assessed. Minimum required sample sizes to achieve 80% and 90% power (α = .05, one-tailed) were calculated for a) the point estimate of the random effects model, and b) the lower bound of the 95% CI, or smallest observed effect size of the included studies if the random effect CIs crossed zero. In addition, the smallest reliably detectable effect for 80% and 90% power (α = .05, one-tailed) based on median sample sizes across included studies was calculated for each outcome to demonstrate the difference between current practice and what is needed to reliably detect effects of simulation training in future studies.

Sample sizes for most included studies were very small. Although observed effects were large in most cases, it is likely that these effects are inflated as a function of sampling error. The small number of studies in each analysis, alongside funnel plot distributions may suggest publication bias. The true effect of simulation training are likely to be over-estimated. One observation from [table 1](#) is that the median sample size for three out of five studies falls below the minimum required sample size to reliably detect the point estimate with a minimum of 80% power. For lower bound estimates, reported median sample sizes fall between 1.78 and 13 times smaller than necessary to reliably detect those effects at 80% power, which rises to between 3.38 and 17.3 times smaller at 90% power, dependent on the outcomes. If publication bias is present, then even the lower bound estimates may be anti-conservative in some cases. Unless prospective studies adopt larger sample sizes, or engage with *a priori* rationale for sample sizes (cf. Lakens, 2022),[46] opportunities for reliably identifying the benefit of simulation training may be missed, alongside opportunities for improving curricula and refining assessment methods.

DISCUSSION

Our aim was to carry out a systematic review of randomised trials to find out if such resource worthy simulation-based education leads to any benefits in actual real life surgical practice. Because of the rapid development and advancement in technology in simulation-based education in the present millennium, we decided to restrict our review to the published literature between 2000 and 2020. We excluded trials where participants were medical students because we wanted to include participants who were committed to a career in surgery and hence had the motivation and desire to improve and do better in their career.

OUTCOME MEASURES

So many different outcome measures including the various scoring systems used by different authors makes it very difficult to carry out a meta-analysis and provide conclusive remarks on overall trends. Evaluation of skill levels and operative proficiency in the real-life operating room appears to be a major challenge in trials of this kind. Of all the outcome measures Global Assessment pf operative performance based on direct observation is said to have superior validity and reliability compared to evaluation with the help of check lists.[47-50]
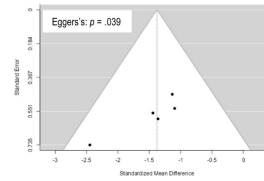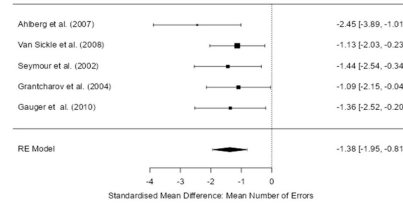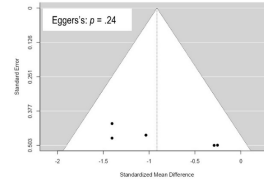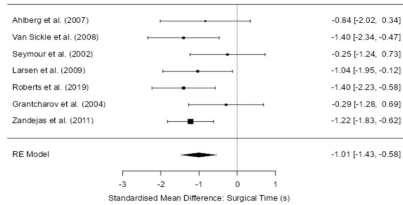
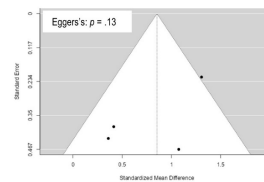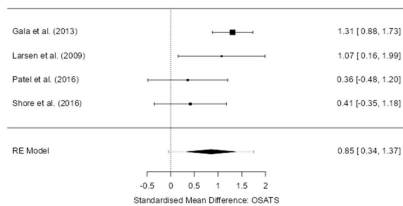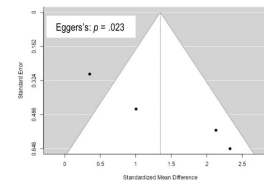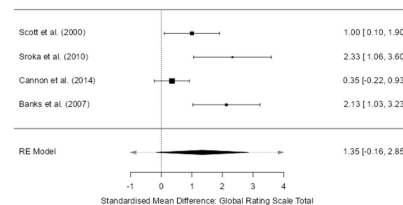**Figure 2.** **Random effects meta-analyses and heterogeneity statistics (left), forest plots (middle), and funnel plots with Egger's test (right).**
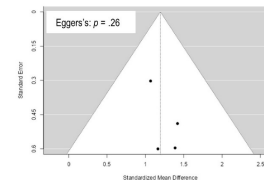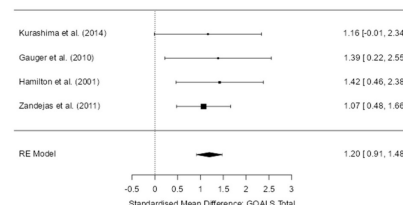
PREDOMINANCE OF TRIALS ON LAPAROSCOPIC PROCEDURE

It is obvious from the above-mentioned results of the included trials that SBT does indeed result in transfer of skills to real life practice in the operating room. Therefore, the efforts and resource spent in providin1g SBE in Surgical discipline appears justifiable.

It may not be altogether surprising for many readers to see that seventeen of the RCT s included in this review involved a laparoscopic surgical procedure.

The remaining two were trials on knee arthroscopy. Throughout the history of surgical training, the focal point of training has been the operating room. Laparoscopic surgery has become the gold standard for many abdominal surgical procedures.[33] As a result, training in laparoscopic

Table 11. Prospective power analyses for all five outcome measures

	Random Effect Estimate [95% CI]	Minimum Per Group Sample Size – Point Estimate			Minimum Per Group Sample Size – Lower bound/Smallest observed			Minimum reliably detectable effect size for median observed sample size		
		Effect Size	80% power	90% power	Effect Size	80% power	90% power	Median group n	80% power	90% power
Error Rates	-1.38 [-1.95, -0.81]	-1.38	$n = 8$	$n = 10$	-0.81	$n = 20$	$n = 27$	8	1.31	1.54
Surgical Time	-1.01 [-1.43, -0.58]	-1.01	$n = 13$	$n = 18$	-0.58	$n = 38$	$n = 52$	11	1.10	1.29
OSATS	0.85 [0.34, 1.37]	0.85	$n = 18$	$n = 25$	0.34	$n = 137$	$n = 183$	13	1.15	1.33
GRS	1.35 [-0.16, 2.85]	1.35	$n = 8$	$n = 11$	0.35*	$n = 130$	$n = 173$	10	1.16	1.36
GOALS	1.20 [0.91, 1.48]	1.20	$n = 10$	$n = 13$	0.91	$n = 16$	$n = 22$	9	1.23	1.44

Notes: All power analyses were conducted one-tailed, with $\alpha = 0.05$. *= Random effects estimate crosses zero, smallest observed point estimate of 0.35 used from Cannon et al., (2014).

surgery has become a common subject of educational research in the field of surgical education. This is mainly due to the fact that learning the skills in laparoscopic surgery is relatively more difficult compared to traditional open surgical procedures because of loss of three-dimensional visualization, lack of tactile feedback and counterintuitive movements of instruments which are often inflexible. Therefore, the apprenticeship model of training doesn't fully fit in to this kind of surgical training, which is best achieved outside the operating room, certainly at least during the majority of learning curve.

In the trials included here, there were wide variation in methods of simulation-based education, methods of intervention and assessment of performance to detect transfer of skills to real life practice.

SIMULATORS

In the seven trials with laparoscopic cholecystectomy, Ahlberg et al,[33] Gauger et al[28] and Hogle et al[32] used Lap-Sim Virtual Reality for training their participants in the SBE group, Grantcharov[9] and Seymour[6] used Minimally Invasive Surgical Trainer – Virtual Reality (MIST-VR), Scott[34] used a video trainer and Sroka[38] used Fundamentals of laparoscopic surgery programme of the American College of Surgeons. For Trials on laparoscopic salpingectomy, four authors used four different kinds of simulators (Larsen,[25] Coleman,[27] Shore,[35] Patel[37]). For laparoscopic bilateral tubal ligation Banks and Gala used a standard laparoscopic simulator produced by Limbs and Things, Bristol, UK.[24,31]

For their trials with knee arthroscopy, Cannon (2014)[39] used ArthroSim Virtual Reality Simulator and Roberts (2019)[40] used a dry knee arthroscopy model along with an American Board approved simulator. These simulators taught the skills that were needed by the novice learners to perform the required surgical procedure. In all the trials SBT was supervised by trainers who set proficiency targets to be achieved before the real-life surgical procedure was performed in the operating room. Simulators alone cannot provide a wholesome rounded training programme. Learners wanting to perform a procedure need to know, what to do and what not to do, how to do it and how to identify when they make an error. That is where lies the role of a good training supervisor who needs to know how the trainee is progressing and where in the learning curve the trainee's ability is positioned and that training includes both psychomotor and cognitive learning. Unless simulators are integrated in an appropriately structured curriculum, their true potential may not be harnessed adequately. Nevertheless, use of various different types of simulators by trainers across the world makes it difficult to structure an agreed curriculum of uniform and satisfactory standard.

ASSESSMENT

In this systematic review we see large variation of methods or tools of assessment even for the same procedure. For laparoscopic cholecystectomy, Ahlberg used mean number of errors, surgical procedural duration and number of conversions from laparoscopic surgery to open procedure.[33] For

the same procedure Gauger (2010) used[28] GOALS and average of total number of errors recorded by each assessor, Grantcharov (2004)[9] used economy of movement, duration of procedure and an error score, Hogle (2009)[32] used a 5-point Likert

GRS which assessed 7 aspects of surgical skills, Scott (2000)[34] and Sroka (2010)[38] used the similar GRS but devised by different authors, Reznick 1997[49] and Vassiliou 2005.[51] Scott assessed the performance of the entire procedure of laparoscopic cholecystectomy.[34] Grantcharov assessed the clipping and cutting portion of the procedure while Seymour assessed only the excision of gallbladder from the liver.[9]

For assessment of performance in knee arthroscopy, Cannon (2014)[39] used procedural checklist, visualization scale, probing scale and global rating, whereas Roberts (2019)[40] used wireless elbow worn motion sensors to objectively assess surgical performance metrics in addition to minor movements, smoothness and time taken.

However, despite these variations, the participants who underwent SBT before performing surgical procedure in real life patients in operating room performed significantly better than their colleagues who did not have SBT (Controls). The SBT group did not always show better performance in all aspects of assessment when compared to their counterpart Control group, but the fact is that the Control group never showed superiority over the SBT group.

Gallagher et al (2005)[52] suggested that measurement of surgical error is the most valuable metrics that SBT can provide in the assessment of competence. Ahlberg et al (2007),[33] Van Sickle et al (2008),[30] Seymour et al (2002)[6] and Grantcharov et al (2004)[9] used error rates or error scores in their assessment of performance.

Amongst the nineteen trials included here, only one trial (Seymour et al 2002) mentioned instances of takeover by the attending surgeon.[6] In this trail there were six such instances when the attending surgeon had to take over.

These were marked as 'errors' and occurred exclusively in the standard programmatic training group (Controls). No statistical significance was documented. These instances indicate a dangerous lack of competence or ability to operate and constitutes great risk to patient safety. Such failures are likely to be rooted in lack of competences which are beyond technical skills alone and may not be mastered by training on simulators alone. It is vitally important that surgeons acquire non-technical skills through structured training process which can then lead to enhanced patient safety.

DURATION OF TRAINING AND TIME FACTOR IN ASSESSMENT

In all the trials the SBT group were trained until they reached predefined targets or achieved proficiency levels set by the trialists. The time taken for such training varied. Assessment of proficiency was tested almost immediately after the completion of training. One author (Gala et al, 2013)[31] reported that they found it difficult to facilitate the completion of their trial by many residents during a single rotational block and also because of the variation in

the number of patients seeking the operation being tested which was sometimes worsened by last minute cancellation of operations. However, they did not find any statistically significant difference between their SBT and Control group. Whether the duration of training or the minimum number of cases or both are important criteria in determining transfer of skills seems to be uncertain although both parameters have its proponents (Casa, 1999; Strum et al, 2008).[53,54]

Brunner et al (2004), suggest that training programmes solely based on duration of training or number of repetitions may be inadequate for acquiring skills because learning curves vary and can be lengthy in novice learners.[55]

Some authors used surgical time or duration of procedure as outcome measures of proficiency.[6,25,30,33,40] But Gauger et al (2010)[28] feels that speed of operation is not an appropriate surrogate for proficiency-based training.

LIMITATIONS

The sample sizes in the included trials were mostly less than 25 participants in each trial. Only 5 of the trials recruited more participants of which 2 trials had 27 participants (Shore et al 2016)[35] and 30 participants (Roberts et al,2019).[40] This limitation was acknowledged by the respective authors and usually attributed to the limited number of post graduate resident trainees in the rotational training programme.

In nine of the nineteen trials, a statistical calculation for sample size and power was undertaken. In four of these nine trials, Gala et al 2013,[31] Kurashima et al 2014,[36] Roberts et al, 2019[40] and Scott et al, 2000,[34] the number of participants who eventually completed the study fell short of the calculated sample size. In the trial by Kurashima et al (2011),[56] the mean GOALS scores which formed the basis of power calculation was not achieved. In the remaining ten trials the authors did not declare any statistical calculation of power of study or sample size.

This objective of this systematic review was to find out skills transfer from the simulation-based training (SBT) to the real-life surgical procedure in operating room and as such trials included in this review were irrespective of patient outcome-based assessment or type of simulation used. Determinants of transfer of skill include design of simulator, functional capability of simulator, design of the training programme, preparation before SBT, nature and type of formative and summative feedback and opportunity for remedial or corrective measures to be taken for any shortcomings detected during SBT. This means that the evidence favouring transfer of skills should not be attributed to simulation alone.

Both the control group as well as the SBT group were undergoing on the job training as part of their residency programme and are likely to have encountered the same operation during their regular resident training. This kind of additional out-of-trial experience would have added to their resultant knowledge and skills which would be wrong to attribute to SBT only as part of the trial. In addition, there is the possibility of assessor bias regarding the abilities of the residents because the recruited residents were already working in the same institutions as part of their ro-

tational training programme. This phenomenon might have affected the evaluation of trial participants and outcome of the trials.

Gala et al (2013)[31] recruited the largest number of participants, that is 102, but unfortunately assessment of skills on laparoscopic bilateral tubal ligation was not a blinded process. The authors admit the unblinded design of the randomization assignment. They did try to mitigate for this by separating the SBT teachers from the surgical proctors and also asked the participants not to disclose their randomization to others. In another trial Hogle et al (2009)[32] assessors evaluated operative skills of participants on elective laparoscopic cholecystectomy in an unblinded manner. Again, the authors admitted to this shortcoming.

The various different simulators, variety of assessment tools used, and different endpoints of training is likely to have led to inconsistencies in assessment making it difficult to make conclusive remarks on the skills achieved at the end of the training and the outcome of assessment.

In many of the trials, assessment and evaluation of competence was undertaken on video recordings of the operative procedure by trial participants. This is likely to have limited the capacity to extensively review errors because the field of view would be restricted to the field of operation and potentially exclude views outside the patient's abdomen. Scott et al (2000)[34] suggest that for assessment of differences in performance, direct observation is superior to video analysis.

Also, we don't know how long the skills acquired from SBT lasts. Assessment of skills and proficiency on real life patients in the trials included here were undertaken within a very short time period of the SBT. Whether long term maintenance of such acquired skills requires regular 'booster doses' of SBT, remains unclear. Hence the transfer of skills from the SBT room to the real-life operating room might be a transient or temporary phenomenon.

CONCLUSION

The ultimate success of SBT depends transfer of skills from the simulated setting to real life operating rooms. The objectives of running a SBT course can be evaluated by application of Kirkpatrick levels 1, 2 and 3, namely: 1) reaction of the learner 2) learning by the learner 3) behavioural change.[47] These Kirkpatrick levels are similar to phase 1 of evaluation strategy suggested in translational science research (TSR) method.[48]

For assessing the utility of SBT courses, Kirkpatrick's levels and Translational Science Research framework are two practically useful frameworks. Their description are as follows.

Kirkpatrick level 1: Learner's reaction to the process of learning. There is no corresponding phase for this stage of learning in TSR framework.

Kirkpatrick level 2: Degree or extent of enhancement of learner's knowledge and skills. TSR phase 1 corresponds to this level and is demonstrated by learning in simulation course.

Kirkpatrick level 3: Skills transferred to capability to perform the practical procedure in real life job or clinical practice. TSR phase 2 equated to this level.

Kirkpatrick level 4: Influence or effect of the SBE course on patient safety. TSR phase 3 is equivalent to this level as it is used to demonstrate whether there was any significant improvement in outcome for patients as a result of the skills acquired in the SBE course.

At present we undertake evaluation of our course at Kirkpatrick levels 1, 2 and partially at level 3.

It may be inappropriate to draw firm and ultimate conclusions on the basis of this systematic review because of methodological inconsistency, differing types of simulation-based training, heterogeneity of outcome measures and possible publication bias for positive or significant trials only, and sample sizes that are likely to be too small to reliably detect realistic effect sizes. Patient outcome was assessed in only one trial and those outcomes were short term.

However, on balance of the available evidence, this review shows that mean error rate was significantly less in the SBT group when compared to the Control group. Mean surgical duration (time) was less in the SBT group when compared to the Control group. Mean OSATS score was higher in the SBT group when compared to the Control group indicating that the SBT group performed better than Controls. GRS score was higher in the SBT group when compared to Control group, suggesting improved skills in SBT group, but this betterment was not statistically significant. Global operative assessment of operative laparoscopic skills (GOALS) score was significantly better in the SBT group suggesting a clear improvement in skills in the SBT group. There appears to be publication bias in estimation of mean error rates and GRS scores.

Larger adequately powered trials should be carried out employing widely available standard simulation-based training, using well defined validated outcome measures, consistent techniques of assessment including assessment of both short- and long-term patient outcomes.

REFERENCES

1. Cameron JL. William Stewart Halsted: Our surgical heritage. *Annals of Surgery*. 1997;225(5):445-458. [doi:10.1097/00000658-199705000-00002](doi:10.1097/00000658-199705000-00002)

2. Bridges M, Diamond DL. The financial impact of teaching surgical residents in the operating room. *Am J Surg*. 1999;177(1):28-32. [doi:10.1016/S0002-9610(98)00289-X](doi:10.1016/S0002-9610(98)00289-X)

3. Scott DJ, Bergen PC, Rege RV, et al. Laparoscopic Training on Bench Models: Better and More Cost Effective than Operating Room Experience? *J Am Coll Surg*. 2000;191(3):272-283. [doi:10.1016/S1072-7515(00)00339-2](doi:10.1016/S1072-7515(00)00339-2)

4. Aggarwal R, Mytton OT, Derbrew M, et al. Training and simulation for patient safety. *Qual Saf Health Care*. 2010;19(Suppl 2):i34-ei43. doi:10.1136/qshc.2009.038562

5. Fletcher JD, Wind AP. Cost considerations in using simulations for medical training. *Military medicine*. 2013;178(10):37-46.

6. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, et al. Virtual reality training improves operating room performance results of a randomized, double-blinded study. *Annals of Surgery*. 2002;236(6):458-463. [doi:10.1097/00000658-200210000-00008](doi:10.1097/00000658-200210000-00008)

7. Hamilton EC, Scott DJ, Fleming JB, Rege RV, Laycock R, Bergen PC, et al. Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills. *Surg Endosc Other Interv Tech*. 2002;182:725-728. [doi:10.1007/s00464-001-8149-z](doi:10.1007/s00464-001-8149-z)

8. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, et al. Proving the value of simulation in laparoscopic surgery. *Annals of Surgery*. 2004;240(8):518-525. [doi:10.1097/01.sla.0000136941.46529.56](doi:10.1097/01.sla.0000136941.46529.56)

9. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg*. 2004;91:146-150. [doi:10.1002/bjs.4407](doi:10.1002/bjs.4407)

10. Munz Y, Kumar BD, Moorthy K, Bann S, Darzi A. Laparoscopic virtual reality and box trainers: Is one superior to the other? *Surg Endosc Other Interv Tech*. 2004;18:485-494.

11. Andreatta PB, Woodrum DT, Birkmeyer JD, Yellamanchilli RK, Doherty GM, Gauger PG, et al. Laparoscopic skills are improved with LapMentorTM training: Results of a randomized, double-blinded study. *Ann Surg*. 2006;243(6):854-860. [doi:10.1097/01.sla.0000219641.79092.e5](doi:10.1097/01.sla.0000219641.79092.e5)

12. Park J, MacRae H, Musselman LJ, Rossos P, Hamstra SJ, Wolman S, et al. Randomized controlled trial of virtual reality simulator training: transfer to live patients. *Am J Surg*. 2007;194:205-211. doi:10.1016/j.amjsurg.2006.11.032

13. Gurusamy KS, Aggarwal R, Palanivelu L, Davidson BR. Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane Database of Systematic Reviews*. Published online January 21, 2009. [doi:10.1002/14651858.CD006575.pub2](doi:10.1002/14651858.CD006575.pub2)

14. Kundhal PS, Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surg Endosc Other Interv Tech*. 2009;23:645-649. [doi:10.1007/s00464-008-0043-5](doi:10.1007/s00464-008-0043-5)

15. Crochet P, Aggarwal R, Dubb SS, Ziprin P, Rajaretnam N, Grantcharov T, et al. Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical skills. *Ann Surg*. 2011;253:1216-1222. [doi:10.1097/SLA.0b013e3182197016](doi:10.1097/SLA.0b013e3182197016)

16. Palter VN, Orzech N, Reznick RK, Grantcharov TP. Validation of a structured training and assessment curriculum for technical skill acquisition in minimally invasive surgery: A randomized controlled trial. *Ann Surg*. 2013;257(16):224-230. [doi:10.1097/SLA.0b013e31827051cd](doi:10.1097/SLA.0b013e31827051cd)

17. Brinkmann C, Fritz M, Pankratius U, Bahde R, Neumann P, Schlueter S, et al. Box- or Virtual-Reality Trainer: Which Tool Results in Better Transfer of Laparoscopic Basic Skills?—A Prospective Randomized Trial. *J Surg Educ*. 2017;74(4):724-735.

18. Kim SC, Fisher JG, Delman KA, Hinman JM, Srinivasan JK. Cadaver-Based Simulation Increases Resident Confidence, Initial Exposure to Fundamental Techniques, and May Augment Operative Autonomy. *Journal of Surgical Education*. 2016;73(6):e33-41. doi:10.1016/j.jsurg.2016.06.014

19. Hooper J, Tsiridis E, Feng JE, Schwarzkopf R, Waren D, Long WJ, et al. Virtual Reality Simulation Facilitates Resident Training in Total Hip Arthroplasty: A Randomized Controlled Trial. *J Arthroplasty*. 2019;34(10):2278-2283. doi:10.1016/j.arth.2019.04.002

20. Lohre R, Bois AJ, Pollock JW, Lapner P, McIlquham K, Athwal GS, et al. Effectiveness of Immersive Virtual Reality on Orthopedic Surgical Skills and Knowledge Acquisition among Senior Surgical Residents: A Randomized Clinical Trial. *JAMA Netw Open*. 2020;3(12):e2031217. doi:10.1001/jamanetworkopen.2020.31217

21. Kohn L, Coorigan J, Donaldson MS, eds. *To Err Is Human: Building a Safer Health System. Summary. To Err Is Human: Building a Safer Health System*. National Academies Press (US); 1999.

22. Pfaff H. Surgical safety and overwork. *British Journal of Surgery*. 2004;91:1533-1535. doi:10.1002/bjs.4829

23. Gurusamy KS, Nagendran M, Toon CD, Davidson BR. Laparoscopic surgical box model training for surgical trainees with limited prior laparoscopic experience. *Cochrane Database of Systematic Reviews*. 2014;(3).

24. Banks EH, Chudnoff S, Karmin I, Wang C, Pardanani S. Does a surgical simulator improve resident operative performance of laparoscopic tubal ligation? *Am J Obstet Gynecol*. 2007;197(24):541.e1-541.e5. doi:10.1016/j.ajog.2007.07.028

25. Larsen CR, Soerensen JL, Grantcharov TP, et al. Effect of virtual reality training on laparoscopic surgery: Randomised controlled trial. *BMJ*. 2009;338(b1802). doi:10.1136/bmj.b1802

26. Zendejas B, Cook DA, Bingener J, et al. Simulation-Based Mastery Learning Improves Patient Outcomes Laparoscopic Inguinal Hernia Repair. A Randomized Controlled Trial. *Ann Surg*. 2011;254:502-511. doi:10.1097/SLA.0b013e31822c6994

27. Coleman RL, Muller CY. Effects of a laboratory-based skills curriculum on laparoscopic proficiency: A randomized trial. *Am J Obstet Gynecol*. 2002;186:836-842. doi:10.1067/mob.2002.121254

28. Gauger PG, Hauge LS, Andreatta PB, et al. Laparoscopic simulation training with proficiency targets improves practice and performance of novice surgeons. *The American Journal of Surgery*. 2010;199:72-80. doi:10.1016/j.amjsurg.2009.07.034

29. Hamilton EC, Scott DJ, Kapoor A, et al. Improving operative performance using a laparoscopic hernia simulator. *Am J Surg*. 2001;182:725-728.

30. Van Sickle KR, Ritter EM, Baghai M, et al. Prospective, Randomized, Double-Blind Trial of Curriculum-Based Training for Intracorporeal Suturing and Knot Tying. *J Am Coll Surg*. 2008;207(4):560-568. doi:10.1016/j.jamcollsurg.2008.05.007

31. Gala R, Orejuela F, Gerten K, et al. Effect of Validated Skills Simulation on Operating Room Performance in Obstetrics and Gynecology Residents. A Randomized Controlled Trial. *Obstet Gynecol*. 2013;121(3):578-584.

32. Hogle NJ, Chang L, Strong VEM, et al. Validation of laparoscopic surgical skills training outside the operating room: a long road. *Surg Endosc*. 2009;23:1476-1482. doi:10.1007/s00464-009-0379-5

33. Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *American Journal of Surgery*. 2007;193:797-804. doi:10.1016/j.amjsurg.2006.06.050

34. Scott DJ, Bergen PC, Rege RV, Dedy NJ, McDermott CD, Lefebvre G. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surgs*. 2000;191(34):272-283.

35. Shore EM, Grantcharov TP, Husslein H, Shirreff L, Dedy NJ, McDermott CD, et al. Validating a standardized laparoscopy curriculum for gynecology residents: a randomized controlled trial. *Am J Obstet Gynecol*. 2016;215:204.e1-11.

36. Kurashima Y, Feldman LS, Kaneva PA, et al. Simulation-based training improves the operative performance of totally extraperitoneal (TEP) laparoscopic inguinal hernia repair: A prospective randomized controlled trial. *Surg Endosc*. 2014;28:783-788.

37. Patel NR, Makai GE, Sloan NL, Della Badia CR. Traditional Versus Simulation Resident Surgical Laparoscopic Salpingectomy Training: A Randomized Controlled Trial. *Journal of Minimally Invasive Gynecology*. 2016;23(3):372-377. doi:10.1016/j.jmig.2015.11.005

38. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM. Fundamentals of Laparoscopic Surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. *Am J Surg*. 2010;199(38):115-120.

39. Cannon WD, Garrett WE, Hunter RE, et al. Improving residency training in arthroscopic knee surgery with use of a virtual-reality simulator: A randomized blinded study. *J Bone Joint Surg - Am Vol.* 2014;96:1798-1806. doi:10.2106/JBJS.N.00058

40. Roberts PG, Alvand A, Gallieri M, Hargrove C, Rees J. Objectively assessing intraoperative arthroscopic skills performance and the transfer of simulation training in knee arthroscopy: a randmised controlled trial. *Arthroscopy.* 2019;35(4):1197-1209. doi:10.1016/j.arthro.2018.11.035

41. Ahmed AA, Muhammad RA. A Beginners Review of Jamovi Statistical Software for Economic Research. *Dutse International Journal of Social and Economic Research.* 2021;6(1):109-118.

42. Knapp G, Hartung J. Improved test for a random effects meta-regression with a single covariate. *Stat Med.* 2003;22(17):2693-2710. [doi:10.1002/sim.1482](doi:10.1002/sim.1482)

43. Morey RD. GitHub - richarddmorey/jpower: functions to compute power for various designs. Accessed January 12, 2023. https://github.com/richarddmorey/jpower

44. Rohatgi A. WebPlotDigitizer (version 4.5). Published 2021. https://automeris.io/WebPlotDigitizer

45. McGrath S, Zhao XF, Steele R, et al. Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. *Statistical Methods in Medical Research.* 2020;29(9):2520-2537.

46. Lakens D. Sample size justification. *Collabra Psychology.* 2022;8(1):1-28. doi:10.1525/collabra.33267

47. Anastakis DJ, Regehr G, Reznick RK, et al. Assessment of technical skills transfer from the bench training model to the human model. *Am J Surg.* 1999;177:167-170. [doi:10.1016/S0002-9610(98)00327-4](doi:10.1016/S0002-9610(98)00327-4)

48. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a Structured Technical Skills Assessment Form. *Am J Surg.* 1994;167:423-427.

49. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative 'bench station' examination. *Am J Surg.* 1996;173:226-230. [doi:10.1016/S0002-9610(97)89597-9](doi:10.1016/S0002-9610(97)89597-9)

50. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(50):273-278. [doi:10.1046/j.1365-2168.1997.02502.x](doi:10.1046/j.1365-2168.1997.02502.x)

51. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190:107-113.

52. Gallagher AG, Ritter M, Champion H, et al. Virtual reality simulation for the operating room. Proficiency based training as a paradigm shift in surgical skills training. *Annals of Surgery.* 2005;241(2):364-372.

53. Cassa OW. Training to competence in gastrointestinal endoscopy: a plea for contnuous measuring of objective end points. *Endoscopy.* 1999;9:751-754. [doi:10.1055/s-1999-148](doi:10.1055/s-1999-148)

54. Strum LP, Windsor JA, Cosman PH, Cregan P, Hewett PJ, Maddern GJ. A systematic review of skills transfer after surgical simulation training. *Annals of Surgery.* 2008;248(2):166-179. [doi:10.1097/SLA.0b013e318176bf24](doi:10.1097/SLA.0b013e318176bf24)

55. Brunner WC, Korndorffer JR, Sierra R, et al. Laparoscopic virtual reality training: Are 30 repetitions enough? *Journal of Surgical Research.* 2004;122:150-156.

56. Kurashima Y, Feldman LS, Al-Sabah S, Kaneva PA, Fried GM, Vassiliou MC. A tool for training and evaluation of laparoscopic inguinal hernia repair: The global operative assessment of laparoscopic skills-groin hernia (GOALS-GH). *Am J Surg.* Published online 2011:54-61. doi:10.1016/j.amjsurg.2010.09.006