Strategies to improve acquisition of technical skill in surgical residents: from screening technical ability at the time of selection to incorporating performance adjuncts during training.

by

Marisa Louridas

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy in Medical Science

Institute of Medical Science
University of Toronto

© Copyright by Marisa Louridas 2016
Strategies to improve acquisition of technical skill in surgical residents: from screening technical ability at the time of selection to incorporating performance adjuncts during training.

Abstract

**Introduction**: Evidence suggests that not all trainees reach technical competence. Therefore the purposes of the included studies were to improve resident selection by investigating screening tools (visual spatial tests (VSTs) and technical tasks (TTs)) that may predict technical ability of incoming trainees, and to determine whether metal practice is beneficial as a performance enhancement strategy during training.

**Methods**: Screening with VSTs as a predictor of laparoscopic ability was evaluated using the PicSOr, cube comparison (CC) and card rotation (CR) tests and correlated to technical performance on the camera navigation (LCN) and
laparoscopic circle cut (LCC) tasks. To screen trainees using TTs, a Delphi of Canadian general surgery (GS) program directors (PD), was performed to gain consensus on the simulated TTs best suited for incoming trainees. K-mean clustering learning curve (LC) analysis was used to determine acquisition of TTs. Next, mental practice was evaluated in a randomized control trial to assess its impact on advanced laparoscopic technical performance.

**Results:** Thirty-seven residents were screened using VSTs. Residents who scored higher on the CC test had more accurate LCN path length ($r_{s(PL)}=-0.36$, $p=0.03$) and angle path ($r_{s(AP)}=-0.426$, $p=0.01$) scores. Eleven of 14 GS PDs participated in the Delphi, and consensus was reached that both basic laparoscopic and open skills would be appropriate for the assessment of TTs. LC analysis of 65 students revealed that 7-15% of trainees did not reach proficiency in laparoscopic skills. These students demonstrated poor innate ability, and remained disadvantaged with inconsistent performance throughout their LC. During training, mental practice significantly improved technical performance ($p =0.003$).

**Conclusion:** LC analysis of simulated technical skills proved more dependable than VSTs to screen for technical ability in novice trainees, while mental practice is an affective adjunct to technical skills performance and would be a beneficial addition to skills training for senior residents.
Acknowledgments

I would like to express my heartfelt gratitude to my program advisory committee Drs. Tulin Cil and Simon Graham, who have provided experienced advice and ongoing support throughout my degree. I would also like to thank my research supervisor Dr. Teodor Grantcharov who has spent countless hours discussing project designs, fine tuning methodology and providing the guidance required to execute novel timely research within the field of surgical education. Through their mentorship I have had the opportunity to publish broadly and present consistently throughout my PhD at both national and international conferences.

This PhD would not have been possible without the Surgeon Scientist Training Program and the Clinical Investigator Program. These programs provide financial support and superb academic resources, to create an environment where clinical residents can dedicate time to pursue advanced degrees in science, forming the foundation for a career as a clinical investigator. Drs. Najma Ahmed, Andy Smith, Carol Swallow, James Rudka, Andrea McCart and Norm Rosenblum, thank you for this opportunity.

Finally I would like to thank my friends, family and colleagues, who have offered words of wisdom, unwavering support, love and balance during the course of this degree. Katy Louridas, Georgia Louridas, Dr. Micahel Zywiwel, Dr. Peter Szasz, Dr. Sandra de Montbrun, Dr. Esther Bonrath, Dr. Nicolas Dedy and Dr. Andras Fecso - thank you!
Contributions

Marisa Louridas independently prepared this thesis and all aspects of the included original research studies from: study design, data collection, analysis and writing. This thesis contains five original manuscripts with Marisa Louridas as the primary author. All contributions by coauthors are described in detail below:

Supervisor – Dr. Teodor Grantcharov – mentorship, guidance for study design, laboratory resources, introductions to collaborators and manuscript/thesis editions.

Thesis committee members – Dr. Tulin Cil and Dr. Simon Graham, study design guidance and thesis preparation

Lauren Quinn – Contributed to the data collection, analysis and preparation of the manuscript in Chapter 3.

Dr. Peter Szasz – assisted in grading quality of studies included in section 2.4 and assisted with study design and participant correspondence in Chapter 4 and trained participants in Chapter 5 while contributing to the preparation and editions of all three manuscripts.

Dr. Sandra de Montbrun – assisted in study design, analysis and manuscript preparation and editions of section 2.4 and Chapter 4.

Dr. Michael Zywiel – assisted in the study design, analysis interpretation and manuscript preparation and editions of Chapter 5.

Dr. Andras Fecso- assisted in preparing the technical skills curriculum trained participants in Chapter 5 while contributing to the preparation and editions of this manuscript

Parisa Lak and Dr. Ayse Bener - experts in data science and machine learning, assisted in the k-means analysis of Chapter 5 and edited the manuscript.

Drs. Esther Bonrath, Dana Sinclair and Nicolas Dedy contributed to the planning, design, execution and manuscript preparation of Chapter 6
# Table of Contents

ACKNOWLEDGMENTS........................................................................................................... IV

CONTRIBUTIONS.................................................................................................................. V

TABLE OF CONTENTS .......................................................................................................... VI

LIST OF TABLES .................................................................................................................... XI

LIST OF FIGURES ................................................................................................................. XIII

LIST OF ABBREVIATIONS .................................................................................................... XV

CHAPTER 1: GENERAL INTRODUCTION ............................................................................. 1
  1.1 Thesis overview ............................................................................................................. 1
  1.2 Hypotheses and Aims ................................................................................................... 2

CHAPTER 2: LITERATURE REVIEW ....................................................................................... 5
  2.1 The current resident selection process ......................................................................... 5
    2.1.1 CaRMS written application ...................................................................................... 6
    2.1.2 National interview process ....................................................................................... 7
    2.1.3 Informal discussion .................................................................................................... 8
    2.1.4 The Canadian national residency match .................................................................... 8
    2.1.5 Evidence for the current selection parameters and their ability to predict resident performance ............................................................................................................................. 9
  2.2 Surgical trainees’ variable success in acquiring technical skill ................................... 13
  2.3 Potential predictors of technical ability for use during the selection process .... 14
    2.3.1 Self-selection as a predictor of technical ability ..................................................... 14
    2.3.2 Surrogate markers as predictors of future technical ability .................................... 16
    2.3.3 The role of simulated technical tasks to inform trainee selection ............................ 41
  2.4 Simulation modalities for technical skill training ......................................................... 43
    2.4.1 Open simulation models .......................................................................................... 44
    2.4.2 Minimally-invasive simulation models .................................................................... 46
2.5 INSTRUMENTS FOR THE ASSESSMENT OF TECHNICAL SKILL ................................................. 49
2.5.1 Task specific checklists ........................................................................................................... 51
2.5.2 Global rating scales ................................................................................................................ 51
2.5.3 Task metrics .......................................................................................................................... 53
2.5.4 Using learning curves to assess technical skill progression ................................................. 54
2.6 TECHNICAL SKILL TRAINING CURRICULA ......................................................................... 55
2.6.1 Structuring curricula for technical skill training ............................................................... 55
2.6.2 Mental practice as an adjunct for technical skill training ................................................ 56

CHAPTER 3: PREDICTIVE VALUE OF BACKGROUND EXPERIENCES AND VISUAL SPATIAL ABILITY TESTING ON LAPAROSCOPIC BASELINE PERFORMANCE AMONG RESIDENTS ENTERING POSTGRADUATE SURGICAL TRAINING ......................................................... 69
3.1 ABSTRACT ................................................................................................................................. 69
3.2 INTRODUCTION ...................................................................................................................... 70
3.3 MATERIALS AND METHODS ............................................................................................... 72
3.3.1 Participants and setting ........................................................................................................ 72
3.3.2 Demographics and participant questionnaire ..................................................................... 72
3.3.3 Visuospatial testing ............................................................................................................. 73
3.3.4 Laparoscopic skills familiarity session ............................................................................... 75
3.3.5 Assessment of laparoscopic baseline skill ........................................................................ 75
3.3.6 Statistical Methods ............................................................................................................. 75
3.4 RESULTS .................................................................................................................................. 76
3.4.1 Demographics ..................................................................................................................... 76
3.4.2 Participant questionnaire .................................................................................................... 79
3.4.3 Correlation of visual spatial skills and laparoscopic baseline performance .................... 79
3.5 DISCUSSION ............................................................................................................................ 80

CHAPTER 4: OPTIMIZING THE SELECTION OF GENERAL SURGERY RESIDENTS - A NATIONAL CONSENSUS ................................................................. 84
4.1 ABSTRACT ................................................................................................................................. 84
4.1.1 Background ......................................................................................................................... 85
4.2 METHODS ............................................................................................................................... 86
4.2.1 Current selection practices ................................................................. 86
4.2.2 Delphi consensus methodology ..........................................................87
4.2.3 Expert panel .........................................................................................87
4.2.4 Anonymity .........................................................................................87
4.2.5 Rounds of questions........................................................................... 88
4.2.6 Consensus .........................................................................................88
4.2.7 Recommendations for incorporation into selection.......................... 88

4.3 RESULTS ..................................................................................................89
4.3.1 Current selection process ................................................................. 89
4.3.2 Delphi consensus methodology .........................................................91

4.4 DISCUSSION ..........................................................................................96

4.5 CONCLUSION ................................................................................--------99

CHAPTER 5: PRACTICE DOES NOT ALWAYS MAKE PERFECT - THE NEED FOR
SELECTION CURRICULA IN MODERN SURGICAL TRAINING .........................101

5.1 ABSTRACT ............................................................................................101
5.2 BACKGROUND .....................................................................................102
5.3 METHODS ...........................................................................................103
5.3.1 Participants ......................................................................................103
5.3.2 Sample Size .....................................................................................104
5.3.3 Demographic questionnaire ...............................................................104
5.3.4 Training curriculum ..........................................................................105
5.3.5 Data analysis and statistical methods ...............................................108

5.4 RESULTS ................................................................................................109
5.4.1 Participants ......................................................................................109
5.4.2 Learning curves patterns for laparoscopic and open skills ................111
5.4.3 Associations between background characteristics and performance clusters........ 116
5.4.4 Individual performance differences between tasks ............................118
5.4.5 Concerning learning curve features for individuals unable to reach proficiency in
laparoscopic tasks ....................................................................................120

5.5 DISCUSSION ..........................................................................................121

5.6 CONCLUSION .......................................................................................125
CHAPTER 6: RANDOMIZED CLINICAL TRIAL TO EVALUATE MENTAL PRACTICE IN ENHANCING ADVANCED LAPAROSCOPIC SURGICAL PERFORMANCE ...............................127

6.1 ABSTRACT ........................................................................................................................................127
6.2 INTRODUCTION ..................................................................................................................................128
6.3 METHODS ............................................................................................................................................129
  6.3.1 Mental practice script development .........................................................................................129
  6.3.2 Randomized trial ......................................................................................................................130
  6.3.3 Statistical analysis ....................................................................................................................133
6.4 RESULTS ................................................................................................................................................133
  6.4.1 Mental practice scripts ..........................................................................................................133
  6.4.2 Randomized trial ....................................................................................................................135
6.5 DISCUSSION .........................................................................................................................................141

CHAPTER 7: GENERAL DISCUSSION ........................................................................................................144

7.1 THESIS SUMMARY ..............................................................................................................................144
7.2 FINDINGS WITH RESPECT TO THE AIMS OF THE THESIS ........................................................145
  7.2.1 To evaluate whether previous surgical experiences, non-surgical experiences and 2D-3D visual spatial tests correlate with laparoscopic skills in the novice surgical trainee 145
  7.2.2 To solicit program directors’ opinions on the proportion of trainees who do not achieve the minimum technical standards expected at the time of graduation ..............................147
  7.2.3 To establish a national consensus on the desired attributes of GS candidates, and the technical skills that would be most indicative of future performance ...........................................148
  7.2.4 To quantify different learning patterns among trainees for both basic and more advanced laparoscopic and open skills ..................................................................................................149
  7.2.5 To determine whether trainees stay within their learning patterns across simulated tasks of varying difficulty (basic and advanced) and type (minimally invasive and open) 150
  7.2.6 To identify a subset of trainees who consistently fail to reach proficiency on simulated tasks and determine the features of their learning curves that separate them from their peers .................................................................................................................................150
7.2.7  To develop a MP script for the performance of an advanced laparoscopic procedure  
      151
7.2.8  To assess the effectiveness of MP on advanced laparoscopic technical skill 
      performance ............................................................................................................. 152
7.2.9  To determine whether MP is associated with differences in stress levels and 
      improvement in non-technical skills in a simulated crisis scenario.......................... 152
7.3    LIMITATIONS .................................................................................................... 153
    7.3.1  Defining cut off scores.................................................................................... 153
    7.3.2  The implementation of global rating scales as a routine assessment method during 
      selection and surgical training .............................................................................. 158
    7.3.3  Adequate sampling........................................................................................... 159

CHAPTER 8: CONCLUSION ............................................................................................... 161

CHAPTER 9: FUTURE DIRECTIONS .............................................................................. 164
    9.1.1  Exploring intraoperative technical skills assessment .................................... 164
    9.1.2  Establish assessment measures of non-technical competencies ...................... 165
    9.1.3  Quantifying mental practice objectively and transferring this techniques into the 
      real operating room ................................................................................................. 165

COPYRIGHT ACKNOWLEDGEMENTS ........................................................................ 167

REFERENCES ............................................................................................................... 169
List of Tables

Table 1: GRADE classification and assessment method of included studies organized by subjective or objective assessment method ................................................................. 25

Table 2: Summary of background characteristics, surgical and non-surgical experiences as predictors of surgical performance collected by participant questionnaires ................. 26

Table 3: A summary of visual spatial tests as predictors of surgical performance ................. 28

Table 4: Summary of dexterity tests as predictors of surgical performance ....................... 32

Table 5: Task specific checklist applicable to both the placement of one interrupted suture and intracorporeal knot tying ......................................................................................... 50

Table 6: Objective Structure of Technical skills - Global Rating Scale (Martin et al., 1997) .... 52

Table 7: Assessment tools of non-technical skills in the operating room .............................. 68

Table 8: Demographic and background characteristics of study participants .................... 77

Table 9: Previous surgical and non-surgical experiences .................................................. 78

Table 10: Correlation of 2D-3D innate ability tests with laparoscopic surgical skill ............. 80

Table 11: Participating Canadian General Surgery Programs ............................................. 90

Table 12: General Surgery program director's responses to clinical knowledge, decision-making and technical skill during selection and at the time of graduation .................. 92

Table 13: Desired candidate attributes for selection into General Surgery ...................... 93

Table 14: Appropriate simulated surgical skills for selection into General Surgery ............ 94

Table 15: Proficiency scores for open tasks ...................................................................... 108

Table 16: Comparing clusters 1-4: start points, end points and repetitions to proficiency .... 114
Table 17: Demographics and non-surgical experiences and their association with performance clusters ........................................................................................................................................ 117

Table 18: Consistency of performance clusters for laparoscopic and open technical skill .... 119

Table 19: Demographics of study participants ........................................................................... 136

Table 20: Result of baseline assessments of technical skill and mental rotation ability .......... 136

Table 21: Technical skill results at baseline and following training .............................................. 140

Table 22: Mental imagery ability at baseline and following training............................................... 141
List of Figures

Figure 1: Flow diagram of search strategy ................................................................. 20

Figure 2: Simulated open technical skills models a. low fidelity synthetic, b. high fidelity
cadaveric porcine, c. high fidelity synthetic (picture provided courtesy of Dr. Sandra de
Montbrun) ......................................................................................................................... 45

Figure 3: (a) the laparoscopic box trainer can be used for both basic and advanced laparoscopic
tasks, using either a low fidelity model such as (b) a plastic penrose drain, or a high fidelity
model such as (c) cadaveric porcine small bowel with attached mesentery ....................... 47

Figure 4: Virtual reality simulator for laparoscopic technical skill training and assessment ...... 49

Figure 5: Task metrics for the laparoscopic knot tie task, using the laparoscopic box trainer .... 53

Figure 6: Examples of expected smoothed learning curves for the performance of (a) a basic task
and (b) a difficult task. Actual performance of an individual learning a basic task may display
additional fluctuations between attempts (c). ...................................................................... 55

Figure 7: Adapted from Lang’s model of input and output variables for emotional imagery .... 59

Figure 8: Mental Practice protocol for surgery, adopted from (S. Arora et al., 2010). ............. 64

Figure 9: Pictorial Surface orientation test (PicSOr) used to assess 2D-3D perception ability. a.
Setup to change between practice and experiment mode. b. rotating arrow oriented to lie 90
degrees to the underlying cube. ......................................................................................... 74

Figure 10: Paper tests used to assess 2D-3D visual spatial ability included the (a) card rotation
test (CR) and (b) cube comparison test (CC) ................................................................. 74

Figure 11: Learning curves for clusters 1 (top performers) to 4 (low performers) on the
laparoscopic (a) peg transfer, (b) circle cut, and (c) intracorporeal knot tie tasks, ordered from
basic to advanced. Proficiency scores are demarcated with a dashed line. ....................... 111
Figure 12: Learning curves for clusters 1 (high performers) to 4 (low performers) on the open (a) one handed tie and (b) laparotomy closure tasks, ordered from basic to advanced. ................... 112

Figure 13: Learning curves for a representative individual from cluster 3 (moderate performer) and from cluster 4 (low performer) on the laparoscopic intracorporeal knot tie, demonstrating variability in performance times and associated normalized scores. ........................................ 121

Figure 14: CONSORT diagram illustrating progress through the phases of the study ........... 135

Figure 15: Comparison of Objective Structured Assessment of Technical Skill (OSATS) change scores between groups. Median (line within box), interquartile range (box), and range (error bars) excluding outliers (circles) are shown. Dotted line indicates baseline performance. P =0 003 (Mann–Whitney U test) ........................................................................................................... 138

Figure 16: Comparison of bariatric Objective Structured Assessment of Technical Skill (BOSATS) change scores between groups. Median (line within box), interquartile range (box), and range (error bars) excluding outliers (circles) are shown. Dotted line indicates baseline performance. P =0 003 (Mann–Whitney U test) ........................................................................ 139

Figure 17: A five-stage model of the mental activities involved in directed skill acquisition (adapted from (Dreyfus & Dreyfus, 1980)) ................................................................. 155

Figure 18: Standard setting using: a. contrasting groups and b. borderline group methodology (images adapted from (de Montbrun, Statterthwaithe, & Grantcharov, 2015)) ................. 156

Figure 19: Standard setting using borderline regression methodology (image adapted from (de Montbrun et al., 2015)) ......................................................................................... 157
List of Abbreviations

TT – technical task

VST – visual spatial test

ACGME - Accreditation Council of Graduate Medical Education

FLS – Fundamentals of laparoscopic surgery

MP – mental practice

MI – mental imagery

MR – Mental rehearsal

CaRMS - Canadian Resident Matching Service

USMLE - United States Medical Licensing Examination

JJ – Jejunojejunostomy

RYGB – Roux-en-y gastric bypass

FLS – Fundamental of Laparoscopic Surgery

MIQ - Mental Imagery Questionnaire

MIQ –RS - Movement Imagery Questionnaire Revised Second version

OSATS – Objective Structure Assessment of Technical Skills

BOSATS – Bariatric Objective Structure Assessment of Technical Skills

STAI – State trait anxiety questionnaire

BP – Blood pressure
HR – heart rate

NOTSS – non-technical skill for surgeons

CC – Cube comparison test

CR – Card rotation test

PicSO or – Pictorial Surface Orientation test

LCC – laparoscopic circle cut

LCN – laparoscopic camera navigation

PGY – post-graduate year

CUSUM - curve cumulative summation test

MS - Motivation-Specific

MG-M - Motivational General-Mastery

MG-A - Motivation General-Arousal

CS - Cognitive Specific

CG - Cognitive General
Chapter 1: General introduction

1.1 Thesis overview

Technical skill ability, development and assessment can be implemented at various stages of surgical training including:

1. At the time of selection – to ensure that students selected to enter surgical residency have the aptitude to reach technical competence during training;

2. During surgical training – to develop adjuncts to technical training for students who have started surgical residency, with the goal of shortening the learning curve in the simulation center to enhance performance in the operating room;

3. Following completion of formal training – to improve already acquired surgical skills and to learn new surgical techniques.

This thesis focuses on the first two training stages with the structure outlined as follows:

Chapter 1: the specific hypothesis and aims of each project.

Chapter 2: the background for the first two training levels mentioned above: selection and in training. A detailed literature review of current selection practices, their ability to predict future resident clinical performance and a systematic review of surrogate markers to predict simulated technical skill. Furthermore, to provide the background to create a technical skills selection curriculum, a review of the pros and cons of different simulated techniques and the available assessment methods for each. Finally, a review on mental practice: the history, hypothesized theories and existing literature in both sports and surgery.
Chapter 3, 4, 5 and 6 are original research manuscripts focusing on technical skills and selection. Chapter 3 is a cross sectional cohort study assessing whether visual spatial tests are able to predict simulated technical performance. Chapter 4 is a Delphi consensus survey on general surgery program directors opinions on the simulation skills most suitable for incoming trainees and chapter 5 is a prospective cohort study, analyzing learning curves using k-means clustering. Lastly, chapter 6 is a randomized control trial of mental practice as a performance enhancement strategy for senior general surgery residents performing advanced laparoscopic skills.

Chapter 7, 8 and 9 include the general discussion, limitations, conclusion and future directions.

1.2 Hypotheses and Aims

At the time of selection

Hypotheses

It is hypothesized that not all surgical trainees will be able to reach technical competence in open and laparoscopic technical skills by the end of training. Therefore, screening incoming medical students for technical aptitude before entry into a surgical program may benefit both the trainee and the training program.

Manuscript 1: Specific surrogate markers that assess visual spatial ability may be appropriate screening tools to predict technical aptitude in laparoscopic skills.

Manuscript 2 and 3: Furthermore, simulated technical tasks may also be used to stratify different learners by assessing their learning curves over multiple repetitions. Within simulated technical tasks, it is hypothesized that medical students will display similar learning curves for disparate
basic laparoscopic and open surgical skills, and that these will be correlated with their potential to reach proficiency in subsequent, more complex technical tasks.

Aims

1: To evaluate whether previous surgical experiences, non-surgical experiences and 2D-3D visual spatial tests correlate with baseline laparoscopic skills in the novice surgical trainee.

2: To identify the current components used in the general surgery selection process at different institutions

3: To solicit program directors’ opinions on the proportion of trainees who do not achieve the minimum technical standards expected at the time of graduation

4: To establish a national consensus on the desired attributes of GS candidates, and the technical skills that would be most indicative of future performance

5: To quantify different learning patterns among trainees for both basic and more advanced laparoscopic and open skills

6: To assess whether background characteristics or experiences explain potential differences in performance.

7: To determine whether trainees stay within their learning patterns across simulated tasks of varying difficulty (basic and advanced) and type (minimally invasive and open)

8: To identify a subset of trainees who consistently fail to reach proficiency on simulated tasks and determine the features of their learning curves that separate them from their peers

During surgical training
Hypotheses

It is hypothesized that mental practice (MP) aimed at teaching the visual and kinesthetic cues for the crucial operative steps of advanced laparoscopic surgery, specifically laparoscopic jejunojununostomy (JJ), will improve surgical performance and decrease stress levels experienced by the surgeon during adverse situations. Furthermore, it is hypothesized that this approach will improve surgeons’ non-technical skills thereby maintaining a competent level of communication, leadership and decision-making during stressful situations.

Aims

1: To develop a MP script for the performance of an advanced laparoscopic procedure

2: To assess the effectiveness of MP on advanced laparoscopic technical skill performance

3: To determine whether MP is associated with differences in stress levels and improvement in non-technical skills in a simulated crisis scenario.
Chapter 2: Literature review

2.1 The current resident selection process

Each year in North America medical and surgical programs endeavor to select trainees who will graduate as competent surgeons. In Canada, the program director is an appointed position and his or her role is to ensure trainees reach competency in their field as defined by the Royal College of Physicians and Surgeons. Furthermore, this individual is responsible to selection, evaluation and remediation of trainees within their program (R. C. o. P. a. Surgeons, 2015).

The selection process in Canada is composed of a structured written application followed by an interview for successful candidates. To submit a structured written application for a surgical residency position in Canada, medical students in their fourth year enter a national match through the Canadian Resident Matching Service (CaRMS). Each surgical program is required to use the CaRMS system, which is an online platform, designed to collect information that will make-up incoming trainees’ applications. Each surgical program specifies which objective and subjective assessment forms the candidate is required to input to complete the written application for their institution. The United States uses a similar process called The National Resident Matching Program (National Resident Matching Program).

The format of the documents inputted into the CaRMS online platform is outlined in detail below (Canadian Resident Matching Service, 2014). Although the each program’s requirements may be slightly different (e.g. requiring 4 reference letters rather than 3), the general format is very similar across the country. In North America the current application form does not contain a separate section to assess technical aptitude. At best, technical performance may be a single checkpoint on the surgical rotation In Training Evaluation Reports (ITER) or a subjective comment within a reference letter(s), however neither of these are standardized or mandatory.
2.1.1 CaRMS written application

1. Curriculum vitae

Verification of Canadian citizenship and medical school attendance.

All candidates require Canadian permanent residency status or citizenship to apply for a post-graduate medical training position in Canada.

2. Clinical clerkship In Training Evaluation Reports (ITERS)

These are institution-specific rating scale assessment tools used by faculty to score the student’s performance within the clinical setting.

3. Examinations

   Medical Council of Canada Examination

   Medical Council of Canada Qualifying Examination (MCCQE) part 1 is a mandatory examination for all medical students before entering supervised practice in a postgraduate training program. In Canada releasing the exam score is optional.

   United States Medical Licensing Examination (USMLE)

   These examinations are mandatory to be licensed medical doctor within the United States of America (USA). Part 1 of 3 is mandatory for application to medical school. These examination results are required for applicants applying for residency training in the USA, however optional for student applying in Canada.

4. Clinical electives

   These are clinical rotations during the final year of medical school in either the discipline or a related discipline to which the candidate is applying.

5. Scholarships and awards
The selection committees generally perceive both scholarly and extracurricular achievements favorably.

6. Research experience and Publications/presentations

An interest in science and research with evidence of productivity also increase the strength of the student’s application.

7. Work experience and Volunteer activity

Programs endeavor to recruit well-rounded applicants therefore extracurricular activities contribute to assessing the applicant’s experiences, interests and commitments outside of their university degree(s).

8. Personal statement

A personalized one page description that explains the applicant’s motivation(s) for applying to the program and their future goals.

9. Reference letters

Each applicant is required to submit 3-5 reference letters written by faculty who can comment on the students’ clinical performance, ability to work within the interdisciplinary healthcare team and overall ability to excel within the program.

Candidates then select the institutions to which they wish to apply and each institution reviews the application as per their own selection criteria. If successful, the candidate is invited for a formal in-person interview.

2.1.2 National interview process

The interview process differs between institutions across the country in terms of the number of interviews required of each candidate, the size of each interview panel, and the interview
questions asked. However, the end goal in most cases is to assess the candidates for characteristics that are not easily obtained from the paper application including: communication skills, enthusiasm for surgery, program fit and interpersonal, problem-solving skills. To do so, each institution hosts an interview day where invited candidates attend in person. However, similar to the paper application, technical performance is not routinely incorporated into the interview process.

2.1.3 Informal discussion

Both medical and surgical programs acknowledge that the combined scores of the written application and the interview do not always adequately assess each applicant holistically. Therefore, informal discussions between the faculty and residents who have worked directly with the students are also part of the selection process. Many of the applicants will travel to the institutions they are most interested in applying to and work clinically with the residents and faculty during a 2-4 week onsite elective. Over this time period, the candidates overall clinical performance may uncover either desirable or undesirable traits that are helpful for the selection committee. Therefore, feedback and input from these encounters are encouraged.

2.1.4 The Canadian national residency match

After the institution has combined the scores of the written application, the interview and the informal discussions, a final rank list of candidates is submitted from each institution to CaRMS. The process of the combining the scores is usually quantitative however the weight of the scores that contribute to the total score are program specific and differ across institutions. Each student also creates their own rank list and both lists are entered into the CaRMS Roth-Peranson algorithm ("CaRMS: The Match Algorithm," 2015). On “national match day”, successful candidates are matched to a program and institution, which they are contractually obligated to attend for their residency training (Canadian Resident Matching Service, 2014).
Therefore, across Canada and the United States, students who successfully enter surgical training do not undergo an assessment of technical aptitude. The present selection process ensures that neither the trainee nor the training program knows whether or not the trainee is adept to acquire the technical skills needed in their surgical training.

2.1.5 Evidence for the current selection parameters and their ability to predict resident performance

Before proposing yet another assessment metric to be added to the already lengthy selection process, a literature review was completed of the current assessments and their predictive potential. Many studies have evaluated the ability of the overall institutional rank score, examination scores, interview scores and reference letters to predict clinical performance during residency. The selection process holds for all trainees entering residency and comprises all program including radiology, pediatrics, surgery and internal medicine. These are outlined below.

2.1.5.1 Overall institutional rank score

The overall rank score for each candidate is the institution’s cumulative score from the written application, interview and information discussions. A measure of recruitment success is the institution’s ability to attract their highest ranked candidates, which are perceived to be the strongest students in the application cohort for their given institution. Intuitively this has been extrapolated to mean that the higher the student is ranked during the selection process, the more successful they will be in the clinical environment. However, this assumed relationship is not supported by the literature. For example, the rank score of radiology residents over 7 years of recruitment at a single program was compared to overall clinical performance scores in their fourth year of residency, with no significant relationship identified between the two measures (Adusumilli et al., 2000). Moreover, of the five candidates that matched during a single recruitment year, the three lowest ranked students were considered the strongest clinical residents by faculty assessment in their fourth year of training (Adusumilli et al., 2000). Similar findings
were reported for a cohort of pediatric residents, where faculty ratings only weakly correlated with residents’ overall rank score during selection (r=0.19, p=0.11) (Borowitz, Saulsby, & Wilson, 2000).

2.1.5.2 Examination scores

Examinations scores are the most objective assessment available to the selection committee. They offer the advantage of comparative assessment of an individual with respect to the national average. However, these scores do not predict future clinical potential. Instead, they predict future performance on upcoming standardized examinations. For instance, successful completion of the USMLE part 1 correlates well with performance on the American Board of Surgery certification examination (de Virgilio, Yaghoubian, Kaji, & et al., 2010). In addition, success on the National Board of Medical Examiners (NBME) part II is predictive of performance on NBME part III (Swanson, Case, & Nungester, 1991). This trend is also true for Medical College Admission Test scores, which predict USMLE scores (Julian, 2005). Common to these examinations is the testing of knowledge acquired through textbook readings, case-based learning and practice questions. It is reasonable that mastery of these skills, coupled with disciplined studying and experience-acquired exammanship may prove beneficial in future written examinations.

However, excellent exam scores do not necessarily render excellent clinicians. Dirscjle et al. examined all 66 resident application files from the University of North Carolina from 1983 to 1997 and demonstrated no significant relationship between resident clinical evaluations by faculty and the Orthopedic In-Training Examination (OITE) and American Orthopaedic Surgery Part I Examination (Dirschl, Campion, & Gilliam, 2006). Woloschuk et al. correlated residency performance in all programs (n=244) to the Medical Council of Canada (MCC) Part 1 examination scores. Resident performance was measured by faculty on a Likert scale from 1 (much weaker than most residents) to 5 (much stronger than most residents) for clinical acumen and human sensitivity. The correlation between the MCC examination and these scores was significant but poor, at 0.17 (clinical acumen) and 0.16 (human sensitivity) (Woloschuk, McLaughlin, & Wright, 2010). Despite the lack of strong correlation between exam scores and clinical performance, when assessing which component of the selection process most influences
overall rank score, performance on the USMLE part 1 examination has been reported as the most significant variable in the United States (Stain et al., 2013). Therefore, high scores on this exam are the most likely predictor of receiving a top 5 rank score at the time of selection, despite the fact that exam scores do not necessarily predict clinical performance (Stain et al., 2013).

In summary, an excellent exam score may predict successful completion of higher-level national examinations, but does not necessarily predict who will become an excellent clinical resident. Although passing the final certification examination for all residents is a necessary millstone before entering independent practice, selecting trainees based on examinations scores will likely have no reliable clinical performance benefit (McGaghie, Cohen, & Wayne, 2011).

2.1.5.3 Interview scores

The purpose of the interview is to assess personality attributes that are not assessed by the written application. However, the capability of interviews to identify residents who are best suited to excel clinically is debated. The predictive utility of this process has been explored in many different medical and surgical disciples with predominantly negative results. Komives et al. compared interview scores from their formalized selection process to in-hospital rating scale performance assessments of 51 residents, in their first two clinical years. The study incorporated residents from a spectrum of different specialties including medicine, surgery and psychiatry. In their institution each applicant underwent two interviews by faculty members, with no significant relationship detected between the their interview scores and in-hospital performance scores (Komives, Weiss, & Rosa, 1984). Similarly, Metro et al. demonstrated no relationship between anesthesia residents’ interview scores and their clinical performance as measured by faculty (Metro, Talarico, Patel, & Wetmore, 2005). In this study, each applicant participated in 4-5 faculty interviews that assessed characteristics such as personality aspects, enthusiasm, assertiveness and maturity, which were averaged to determine the final interview score. In contrast, Borowitz et al. found a weak correlation between pediatric residents’ interview scores and faculty ratings within the clinical setting ($r = 0.27$; p=0.02) (Borowitz et al., 2000). Given the poor associations between interview scores and clinical performance, it has been proposed that
the in-person interview process may be more beneficial in promoting the program to the applicant, rather than predicting future clinical performance (Dubovsky et al., 2008)

2.1.5.4 Reference letters

Reference letters are considered an important component of the application package, and are used to assess the candidate’s performance within the clinical setting as perceived by their supervising faculty. Reference letters, however, have also failed to demonstrate predictive value in terms of future clinical performance across a range of different medical and surgical specialties. When reference letter scores were correlated to clinical performance after the first and third year of post-graduate training in internal medicine, only weak Pearson correlations of \( r = 0.25 \) and \( r = 0.12 \) was found, respectively (Curry, Yarnold, Bryant, Martin, & Hughes, 1988). Similarly, reference letter scores for general surgery residents were found to be weakly correlated with a range of clinical competencies including patient care (\( r=0.35 \)), communication (\( r=0.26 \)) and professionalism (\( r = 0.15 \))(Brothers & Wetherholt, 2007). Furthermore, reference letter scores of obstetrics and gynecology residents also showed no significant relationship with faculty-assessed clinical performance in the categories of clinical judgment and acumen, patient rapport, surgical ability and work ethic (J. G. Bell, Kanellitsas, & Shaffer, 2002).

One potential reason for the poor predictive value of reference letters is reported by Dirschl et al. who determined that the interpretation of the reference letters differed significantly between readers, calling into question the reliability of these letters for assessment (Dirschl & Adams, 2000). Overall, although the use of reference letters is deeply rooted in the traditional application process, these letters have not proven to be predictive of future clinical performance.

2.1.5.5 Summary of the current selection parameters and their ability to predict resident performance

Many of the existing elements of the selection process do no adequately predict future clinical performance, however they continue to persist due to their historic presence and lack of new
predictive assessment tools to replace them. Although many areas of the selection process require study to improve their predictive value, I have chosen to add to the existing literature in this field by focusing on technical skill. Technical skill can be objectively measured and is currently not assessed within the realm of selection. Operating on patients and reaching technical competence is an essential component of surgical practice and one of the key elements that differentiates a surgeon from a medical doctor. Incorporating technical skill into the selection may contribute to further ensuring that the candidates selected for surgical training are best suited for this area of medicine.

2.2 Surgical trainees’ variable success in acquiring technical skill

There is increasing objective evidence that even with continued practice, not all trainees are able to achieve technical competence. Cuschieri et al. performed a longitudinal study of surgical trainees between 1972 and 2002, and found that approximately 5-10% of residents did not reach technical proficiency after completing 5 years of surgical training (Cuschieri, 2003). Unfortunately, the authors of this study did not define ‘technical competence’ nor did they describe how the residents were assessed for proficiency, limiting interpretation of their findings. This phenomenon has been noted in the simulation lab setting as well. Grantcharov et al. examined the learning curves for 37 trainees performing 6 distinct tasks on the Minimally Invasive Surgical Trainer – Virtual Reality (MIST-VR). The authors found that after 10 trials, 8.1% of residents did not show any skill improvement (Grantcharov & Funch-Jensen, 2009). Similarly, Schijven et al. reported that 20% of residents did not reach proficiency on the laparoscopic clip and cut task after 30 trials (M. P. Schijven & Jakimowicz, 2004). With respect to trainees in orthopedic surgery, Alvand et al. reported that after 30 repetitions, 35% of trainees did not reach competence in shoulder arthroscopy, and 25% did not reach competence for knee arthroscopy, with the latter commonly considered to be less technically demanding (Abtin Alvand, Sunil Auplish, Harinderjit Gill, & Jonathan Rees, 2011). Although mostly limited to the simulation environment, these results suggest that approximately 5-35% of the surgical resident training pool is at risk for not reaching technical competence, even with practice.

To the best of my knowledge, no studies to date have investigated the implications of a subset of
surgical residents failing to reach surgical proficiency by the end of training. However, it is reasonable to expect that these individuals absorb greater program resources, and engender frustration on the part of surgeon-teachers and co-residents as a result of being perceived as difficult to teach. From the resident perspective, struggling to gain technical proficiency may increase frustration, promote a sense of inadequacy, and increase the likelihood of transferring to a non-surgical program or dropping out of clinical training all together. Finally, and perhaps most important, recent evidence suggests that poor technical skill in staff surgeons can increase complication rates (Birkmeyer et al., 2013), potentially resulting in adverse consequences in terms of the outcomes of patient care. Therefore, there are several potential advantages to identifying trainees that will be unable to reach technical proficiency, prior to them embarking on a surgical residency.

Adding a technical aptitude assessment test to the existing selection process may be a feasible approach to screening for incoming trainees that may have difficulties in the operating room. The United Kingdom, Ireland and Australasia have all reported ongoing use of technical aptitude testing at the time of selection for some of their surgical programs. These countries emphasize the importance of testing technical potential rather than learnt skills for incoming trainees using either surrogate markers or simulated technical tasks (Louridas, Sazsz, de Montbrun, Harris, & Grantcharov, 2016). Given the current North American selection process described in section 2.1, adding a surrogate technical aptitude test to the interview process would be a feasible approach. To this end, a systematic review of the published literature was performed to identify surrogate markers including personal characteristics and cognitive tests that could potentially be used as part of the selection criteria for applicants to surgical training programs, and that may predict innate technical skill ability.

2.3 Potential predictors of technical ability for use during the selection process

2.3.1 Self-selection as a predictor of technical ability

For a long time, surgical disciplines have thought of themselves as distinct from medical
specialties. Specifically, surgeons take pride in the ability to work well with their hands, often considering themselves ‘doers’ rather than just ‘thinkers.’ Intuitively, one might expect that medical students recognize this characteristic of these disciplines, and studies have shown that students that apply to surgical specialties have a higher self-perceived confidence in dexterity and their ability to ‘work well with their hands’ as compared to students entering other medical specialties (Van Hove et al., 2008). Unfortunately, studies have shown that there is no correlation between students’ subjective self-assessment, and their objective scores on dexterity tests and simulated surgical task performance. Harris et al. had forty-eight trainees in surgery, psychiatry, anesthesiology and medicine undergo objective testing of manual dexterity and hand eye coordination. The authors found no difference in performance between surgical and no-surgical trainees (Harris, Herbert, & Steele, 1994). With respect to performance on simulated surgical tasks, Panait et al. compared basic virtual reality (VR) skills in students entering surgical training to those displayed by residents after a year of internal medicine residency, and found that the internists performed better on three out of four VR tasks (L. Panait et al., 2011). Cope et al. assessed 22 interns, where seven of 10 interns interested in surgery rated themselves as naturally dexterous and only 2 of 12 interns interested in non-surgical disciplines felt they had this ability. However, no significant differences in performance of basic VR skills tasks were identified between the groups, suggesting that higher self-perceived natural dexterity does not confer any objective advantage in technical skill (Cope & Fenton-Lee, 2008). Given these findings, self-selection cannot be relied upon to ensure that surgical trainees have a high potential for technical skill performance.

The differences between self-perceived and objective technical aptitudes may be explained by the relative inexperience of incoming trainees. Even by the end of medical school, students have typically had limited opportunity to practice surgical skills in the operating room, and are unlikely to have obtained an objective evaluation of their technical ability that would allow them to make an informed decision concerning their skills. Therefore, it has become clear that surgical programs cannot rely on students’ perceptions of their technical skill as a surrogate marker for their future technical performance during training. Instead, if technical ability is going to be incorporated into the selection process, then objective assessment of this domain will be essential.
2.3.2 Surrogate markers as predictors of future technical ability

This portion of this chapter has been previously published as Louridas M, Szasz P, de Montbrun S, Harris KA, Grantcharov T. Can we predict technical aptitude? A systematic review. *Ann Surg.* 2015 June 15

2.3.2.1 Background

Becoming a surgeon requires prolonged trainee and faculty commitment, as well as considerable fiscal resources. Surgical residency generally involves 5 to 8 years of specialized training and during these years, trainees work within a grueling, fast-paced, high-stakes environment, which often, results in countless personal sacrifices and high burnout rates (M. Arora, Diwan, & Harris, 2013; Eddlaman, Aoun, & Batjer, 2013; Franke et al., 2013). During this time, residents rely on the teaching and mentorship of surgical faculty to direct and foster the acquisition of clinical knowledge, technical skills, and surgical judgment (Sanfey, Hollands, & Gantt, 2013; Straus, Johnson, Marquez, & Feldman, 2013). In order to fill this role, surgeon educators dedicate both time and energy to deliver formal and informal teaching and technical skill training, and through this impart operative and non-operative surgical judgment to the next generation of surgeons. Furthermore, although outside of the operating room residents improve the efficiency of patient care and decrease length of stay, they significantly increase operative time and the cost of surgical procedures, which impacts both the surgeon educator and the health care system as a whole (Babineau et al., 2004; Chamberlain, Patil, Minja, & Kordears, 2012; Harrington et al., 2007; Offner, Hawkes, Madayag, Seale, & Maines, 2003; Sasor, Flores, Wooden, & Tholpady, 2013). The cost of operating room time in the United States is reported to be approximately $900-1200 per hour, (Macario, 2010) resulting in an estimated cumulative nationwide annual cost burden of $53 million attributed to the extra operating room time used to teach residents each year (Bridges & Diamond, 1999). Most surgical residents, surgical faculty and economists would agree that surgical training is worth the sacrifices as long as it produces safe, competent surgeons who are ready to enter independent practice. To ensure this end goal is reached, surgical programs rely on a structured selection process intended to identify candidates who are most likely to succeed.
The selection processes differ widely between countries and institutions (Accreditation Council for Graduate Medical Education, 2013; CaRMS, 2016; R. A. C. o. Surgeons). However, common to all programs is the desire to select a strong cohort of professionals who are able to learn quickly, work well within the healthcare team, make safe and appropriate clinical decisions, and have the ability to learn the necessary technical skills to operate independently. While technical skill is not commonly a part of the selection process in North America, growing evidence suggests that adding a technical component to the existing selection process should be considered (Cushchieri, 2003; Grantcharov & Funch-Jensen, 2009; Mattar et al., 2013; M. P. Schijven & Jakimowicz, 2004). This added element is pertinent to the modern day trainee who is required to meet technical competency despite restricted resident work hours, introduction of more complex surgical procedures and more diverse application of difficult minimally invasive techniques (Biondi et al., 2013; Khatuja et al., 2014; Levine & Spang, 2014; Richards et al., 2015).

Increasing evidence demonstrates that even with continued practice, not all surgical trainees will achieve technical competence in the operating room by the end of training. Technical competence has been defined as the ability to complete tasks or procedures safely and independently (Szasz, Louridas, Harris, Aggarwal, & Grantcharov, 2015). A single longitudinal study suggested that 5 to 10% of trainees do not reach technical proficiency after completing a 5 year training program (Cushchieri, 2003). Furthermore, a North American survey of fellowship program directors showed that 21% of fellows were unprepared for the operating room with 66% unable to operative independently for more than 30 consecutive minutes (Mattar et al., 2013). In the simulation setting, a number of studies suggest that somewhere between 8.1 to 20% of residents do not reach competence despite ongoing practice of the simulated tasks (Abtin Alvand, Sunil Auplish, Harinderjit Gill, & Jonathan Rees, 2011; Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). This evidence suggests that a proportion of the resident training pool is at risk for not reaching technical competence. Given the individual and faculty commitment as well as the fiscal resources required to train residents, adding a measure of technical skill to the selection process to help identify these individuals and direct them early on to other medical specialties. This approach will benefit the trainees, the educational system and the public (Birkmeyer et al., 2013).
To this end, a systematic review of the published literature was performed to identify trainee background characteristics, as well as cognitive tests (including visual spatial, psychomotor and human basic performance tests) that potentially predict innate technical ability, and therefore could be used to supplement existing surgical residency selection criteria.

2.3.2.2 Methods

Search strategy and criteria

A systematic review was conducted of studies that have evaluated associations between surrogate markers of innate abilities in surgical trainees, and whether these abilities correlate with technical performance, either in the simulation setting or in the operating room. The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement was used to guide the structure of the review (Moher, Liberati, Tetzlaff, & Altman, 2010). One reviewer (M.L.) conducted the search using the online MEDLINE, PsycINFO and Embase online databases from 1946 to August 30, 2013, with the assistance of a full-time medical librarian from St. Michael’s Hospital in Toronto, Canada. The search was performed using the following four sets of medical subject heading (MeSH) terms and keywords combined with the AND function: ‘Surgical Procedure’, Operative/*, ‘surgery or surgeries or surgical or surgically’, ‘Specialties, Surgical’, neurosurg*, ‘exp Laparoscopy’, ‘Laparoscopes’, ‘laparoscopy*’, ‘minimal* adj2 surg*’, operate*or operative* AND ‘exp Education, Medical/*, resident* or residency or residencies’, ‘intern or interns or interne or internes’, ‘trainee*’, ‘registrar*’, ‘house officer*’, ‘candidate*’, ‘applicant*’, ‘student*’, ‘novices’ AND ‘exp Psychomotor Performance’, ‘exp Space Perception/*’, ‘exp Suture Techniques’, ‘skill or skills’, ‘performance’, ‘ability or abilities’, technical’, ‘technique*’ AND ‘School Admission Criteria’, ‘Self-evaluation Program’, ‘Personnel Selection’, ‘exp Psychological Tests’, ‘Aptitude/*’, ‘select*’, ‘preselect*’, ‘aptitude*’, ‘screen*’, ‘accept*’, ‘intake*’, ‘predict*’, ‘determinant*’, ‘criterion or criteria’. A manual cross-reference search of the bibliographies of retrieved articles was performed to identify any additional studies that were relevant to this review.

Eligibility criteria
All original studies that explored the relationship between technical performance in the simulation setting or the operating room, and innate abilities, background characteristics, and/or previous operative and non-operative experiences, were considered eligible for inclusion. Studies that evaluated medical students or surgical trainees (i.e. those that had not yet completed surgical specialty training) were included involving any of the surgical disciplines. No restrictions were placed based on the type of technical task evaluated (e.g. open, laparoscopic or endoscopic). Eligibility was limited to those studies with published abstracts or full text manuscripts available in English. Review articles, expert opinions, case reports and editorials were excluded.

**Data extraction**

The following data were extracted: 1) study features including year of publication, study design, duration and statistical analyses, 2) details of study population including demographics and sample size, 3) surrogate predictors of technical skill and assessment including participant characteristics, visual spatial ability, psychomotor ability (excluding surgical simulation tasks because surrogates designed to simulate operative movements were the focus of this review), and depth perception, 4) type(s) of surgical setting(s) used for assessment, such as virtual reality simulators, box trainers, porcine models or patient operations 5) technical skills assessment measures such as global rating scales, procedures specific checklists, time, or computer generated outputs including path length, path angle and error scores.

**Assessment of methodological quality**

The quality of each selected article was individually assessed by two reviewers (M.L. and P.S.) using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008). Using this framework, quality was assessed by initially stratifying studies by design (randomized trial versus observational study), followed by ranking the evidence up or down based on five defined categories including: limitations, inconsistency, indirectness, imprecision and publication bias.
(Guyatt G. et al., 2011). Any discrepancies in quality were resolved by consensus discussion between the reviewers.

Figure 1: Flow diagram of search strategy

2.3.2.3 Results

In total 8035 citations were initially identified. After sequentially screening by title and abstract, 118 studies remained potentially eligible for inclusion. Full text versions of these studies were subsequently reviewed, with 52 included in the final review (Figure 1). Of the 52 studies, 2 were randomized controlled trials (Brandt & Davies, 2006; Hedman et al., 2006), 1 was a pre-post study (Dashfield, Lambert, Campbell, & Wilkins, 2001), 21 were cross-sectional studies (Dimitriou, Nightingale, Khazali, Hatzigeorgiades, & Prendiville, 2009; Lars Enochsson et al., 2004; L. Enochsson et al., 2006; Gettman et al., 2003; Gibbons, Baker, & Skinner, 1986; Grober et al., 2003; Hassan et al., 2007; Hedman, Klingberg, Enochsson, Kjellin, & Fellander-Tsai, 2007; Hislop et al., 2006; Johnson et al., 2004; Macmillan & Cuschieri, 1999; Madan et al., 2005; Madan, Harper, Frantzides, & Tichansky, 2008; Nomura et al., 2008; Nugent et al., 2012; Shah, Buckley, Frisby, & Darzi, 2003; Shah, Paul, et al., 2003; Steele, Walder, & Herbert, 1992; Tang
et al., 2014; Tangchitnob, Solnik, Saad, Rad, & Ogunyemi, 2011) and 28 were prospective cohort studies (Buckley et al., 2013; Buckley et al., 2014; A. G. Gallagher, Cowie, Crothers, Jordan-Black, & Satava, 2003; Groenier, Schraagen, Miedema, & Broeders, 2014; Hedman et al., 2006; Hoffer & Hsu, 1990; M. Keehner, Lippa, Montello, Tendick, & Hegarty, 2006; M. M. Keehner et al., 2004; Kolozsvári et al., 2011; Masud, Undre, & Darzi, 2012; Marlies P. Schijven, Jakimowicz, & Carter, 2004; Schueneman, Pickleman, Hesslein, & Freeark, 1984; Van Herzeele et al., 2010; Van Hove et al., 2008; Wanzel, Hamstra, Anastakis, Matsumoto, & Cusimano, 2002; Wanzel et al., 2003; White & Welch, 2012). When assessed using the GRADE system, the 2 included RCTs were rated as high quality evidence, 31 studies as low quality evidence, and the remaining 20 as very low quality evidence. The most common reason for downgrading the quality score was a lack of blinding of assessors when using a subjective assessment instrument (20 of 52 studies). A detailed summary of the quality assessment organized by subjective or objective assessment method can be found in Table 1.

Potential predictors of technical ability identified in these studies can be divided into: 1) information generated from participant questionnaires such as background characteristics, non-surgical experiences, and surgical experiences (Table 2); and 2) validated cognitive tests designed to test innate visual spatial ability (Table 3), dexterity (Table 4), human basic performance resources, and other related characteristics (i.e. depth perception, working memory).

### Participant questionnaires

A total of 14 studies attempted to predict surgical performance based on responses to participant questionnaires. Of the 23 potential predictors studied, only video gaming consistently showed a significant correlation with initial technical skill. However, Paschold and Dimitriou both found that gamers lost their initial technical advantage once non-gamers were given the opportunity to practice (Dimitriou et al., 2009; Paschold et al., 2011). All the other characteristics collected by participant questionnaires (e.g. age, handedness, experience typing, sports etc.) failed to consistently predict technical performance (Table 2) (Banerjee, Cosentino, Hatzmann, & Noe, 2010b).
Cognitive tests

Four different categories of cognitive tests, aimed at testing innate abilities perceived to be important in acquiring surgical skill, were identified: 1) visual spatial, 2) dexterity, 3) human basic performance resources and 4) other.

Visual spatial

A total of 38 studies using twenty-five different visual spatial tests (VST) were identified that have been evaluated in terms of their ability to predict technical performance. Of the 25 VSTs, 2 have repeatedly shown a positive relationship with technical performance. Specifically: the PicSOR test of perceptual skill (5 of 8 studies) (Buckley et al., 2013; Buckley et al., 2014; A. G. Gallagher et al., 2003; M. Keehner et al., 2006; Kolozsvari et al., 2011; McClusky, Ritter, Lederman, Gallagher, & Smith, 2005; E. Matt Ritter, McClusky, Gallagher, Enochsson, & Smith, 2006; D. Stefanidis, Korndorffer Jr, et al., 2006), and the mental rotation test (MRT) (6 of 9 studies) (Brandt & Wright, 2005; Deary, Graham, & Maran, 1992; Groenier et al., 2014; Hedman et al., 2007; M. Keehner et al., 2006; Luursema, Buzink, Verwey, & Jakimowicz, 2010; D. Risucci, Geiss, Gellman, Pinard, & Rosser, 2001; Wanzel et al., 2002; Wanzel et al., 2003). When stratified by technical skill category, PicSOR demonstrated a positive relationship for laparoscopic skills learned in the box trainer and virtual reality simulator.(Buckley et al., 2013; Buckley et al., 2014; A. G. Gallagher et al., 2003; Kolozsvari et al., 2011; McClusky et al., 2005) In contrast, MRT has demonstrated a positive relationship with open surgical skills.(Wanzel et al., 2002; Wanzel et al., 2003) All other visual spatial predictors were either only evaluated in a single study, or failed to show a relationship with technical performance in the majority of studies (Table 3) (Buckley et al., 2013; Buckley et al., 2014; Deary et al., 1992; Dimitriou et al., 2009; L. Enochsson et al., 2006; Groenier et al., 2014; M. Keehner et al., 2006; M. M. Keehner et al., 2004; Kolozsvari et al., 2011; Luursema et al., 2010; McClusky et al., 2005; Murdoch, Bainbridge, Fisher, & Webster, 1994; Neumann et al., 2005; Nugent et al., 2012; D. A. Risucci, 2002; E. Matt Ritter et al., 2006; Marlies P. Schijven et al., 2004; Schueneman, Pickleman, & Freeark, 1985; Schueneman et al., 1984; Steele et al., 1992; D. Stefanidis, Korndorffer Jr, et al., 2006; Tang et al., 2014; Van Herzeele et al., 2010; Wanzel et al., 2002).
**Dexterity tests**

A total of 19 studies using twenty different surrogate tests of dexterity were identified. An adaptive tracking task (ADTRACK2) was significantly associated with technical performance in 2 of 2 studies, one testing performance on the reef knot and endoscopic sinus surgery (Dashfield et al., 2001; Dashfield & Smith, 1998). The grooved pegboard test was significantly associated in 4 of 5 studies, however, in these studies, the significant association was either limited to the initial trial of the task, or to a single task sub-score in the areas of laparoscopy and endoscopy (Nugent et al., 2012; D. Stefanidis, Korndorffer Jr, et al., 2006; Van Herzeele et al., 2010). The seventeen remaining dexterity tests were either only evaluated by single study or failed to find a relationship with technical performance (Table 4) (Buckley et al., 2013; Buckley et al., 2014; Hoffer & Hsu, 1990; Macmillan & Cuschieri, 1999; Masud et al., 2012; Murdoch et al., 1994; Neumann et al., 2005; Nugent et al., 2012; Marlies P. Schijven et al., 2004; Schueneman et al., 1985; Schueneman et al., 1984; Steele et al., 1992; D. Stefanidis, Korndorffer Jr, et al., 2006; Van Herzeele et al., 2010; Wanzel et al., 2003).

**Human basic performance resources**

Three studies evaluated association between human basic performance resources (BPRs) and technical performance. BPRs are a group of simple tests including: simple visual-hand response speed, visual information processing speed, upper extremity neuromotor channel capacity, upper extremity steadiness and grip strength. Overall, BPRs accurately predicted technical performance in 62 to 75% of cases. They over predicted performance in 14 to 17% of cases, and under predicted performance in 18 to 21% of cases in the areas of laparoscopic and endoscopy (Gettman et al., 2003; Johnson et al., 2004; Matsumoto et al., 2006).
Other innate abilities

Twelve additional tests of innate ability, which did not fit into the aforementioned categories, were identified. Two of these demonstrated a positive association with technical performance, namely: tonic accommodation, defined as a stable parameter that the eye adopts in the absence of stimulation; and abstract reasoning, which investigates an individual’s non-verbal reasoning and is related to intelligence quotient (M. Keehner et al., 2006; Marlies P. Schijven et al., 2004; Shah, Buckley, et al., 2003). Tonic accommodation was significantly correlated with the error score of the right hand on a virtual reality simulator. However, after the 5th repetition, the relationship was no longer significant (Shah, Paul, et al., 2003). Abstract reasoning was shown to have a positive correlation with laparoscopic cholecystectomy performance on a virtual reality simulator (Marlies P. Schijven et al., 2004). The remaining 10 tests failed to demonstrate an association with technical performance. These included: information management (Dashfield & Smith, 1998), stereoscopic vision (Deary et al., 1992), verbal reasoning (Groenier et al., 2014), working memory (Hedman et al., 2006), flow and mental strain (Hedman et al., 2006), organizational planning (Van Herzeele et al., 2010), memory test (Hedman et al., 2006), flexibility of closure (Luursema et al., 2010), personality test (Neumann et al., 2005) and vigilance endurance test (Neumann et al., 2005).
Table 1: GRADE classification and assessment method of included studies organized by subjective or objective assessment method

<table>
<thead>
<tr>
<th>Study design</th>
<th>Limitations</th>
<th>Precision</th>
<th>Consistency</th>
<th>Directness</th>
<th>Publication bias</th>
<th>Grade</th>
<th>*Studies n=52</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 RCT</td>
<td>not serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>high</td>
<td>31,32</td>
</tr>
<tr>
<td>22 Obs</td>
<td>not serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>low</td>
<td>32,46,67,69,77,73,82,51,52,80,74,54,34,72,37,58,60</td>
</tr>
<tr>
<td>5 Obs</td>
<td>not serious</td>
<td>imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>very low</td>
<td>64,44,59,70,35</td>
</tr>
<tr>
<td>Subjective assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Obs</td>
<td>not serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>low</td>
<td>43,53</td>
</tr>
<tr>
<td>13 Obs</td>
<td>serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>very low</td>
<td>42,33,75,38,41,66,47,68,48,84,81,55,79</td>
</tr>
<tr>
<td>2 Obs</td>
<td>not serious</td>
<td>imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>very low</td>
<td>36,56</td>
</tr>
<tr>
<td>Combination of objective and subjective assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Obs</td>
<td>not serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>low</td>
<td>61,62</td>
</tr>
<tr>
<td>4 Obs</td>
<td>serious</td>
<td>no imprecision</td>
<td>Consistent</td>
<td>direct</td>
<td>not detected</td>
<td>very low</td>
<td>49,50,65,57</td>
</tr>
</tbody>
</table>

RCT - Randomized controlled trial; Obs - Observational study (cross-sectional, cohort, case-series, pre-post quasi-experimental); *Studies - references within the body of the manuscript.
Table 2: Summary of background characteristics, surgical and non-surgical experiences as predictors of surgical performance collected by participant questionnaires

<table>
<thead>
<tr>
<th>Count</th>
<th>Potential predictors (n=23)</th>
<th>Number of studies</th>
<th>Total number of participants (number of participants in significant studies)</th>
<th>Number of studies reporting a significant association (percent; weighted percent)</th>
<th>Number of studies that failed to identify a significant association (percent; weighted percent)</th>
<th>*Studies n=14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Background characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Gender</td>
<td>9</td>
<td>493 (272)</td>
<td>3 (33.3; 55.2)</td>
<td>6 (66.6; 44.8)</td>
<td>71,49,60,38,58,79,74,72,69</td>
</tr>
<tr>
<td>2</td>
<td>Handedness</td>
<td>6</td>
<td>651 (0)</td>
<td>0</td>
<td>6 (100.0; 100.0)</td>
<td>71,50,70,60,38,79</td>
</tr>
<tr>
<td>3</td>
<td>Age</td>
<td>5</td>
<td>548 (35)</td>
<td>1 (20.0; 6.4)</td>
<td>3 (80.0; 93.6)</td>
<td>71,70,60,79,72</td>
</tr>
<tr>
<td>5</td>
<td>Surgical career aspirations</td>
<td>3</td>
<td>401 (32)</td>
<td>1 (33.3; 8.0)</td>
<td>(66.6; 92.0)</td>
<td>70,69,51</td>
</tr>
<tr>
<td>6</td>
<td>Self-reported motor skills</td>
<td>3</td>
<td>390 (43)</td>
<td>1 (33.3; 11.0)</td>
<td>1 (66.6; 89.0)</td>
<td>72,70,51</td>
</tr>
<tr>
<td>7</td>
<td>Glove size</td>
<td>3</td>
<td>57 (11)</td>
<td>1 (33.3; 19.3)</td>
<td>1 (66.6; 80.7)</td>
<td>71,74,72</td>
</tr>
<tr>
<td>9</td>
<td>Weight</td>
<td>2</td>
<td>46 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>71,72</td>
</tr>
<tr>
<td>10</td>
<td>Vision</td>
<td>2</td>
<td>36 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>71,74</td>
</tr>
<tr>
<td>11</td>
<td>Height</td>
<td>2</td>
<td>46 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>71,72</td>
</tr>
<tr>
<td>12</td>
<td>Completion of surgical internship</td>
<td>1</td>
<td>326(0)</td>
<td>0</td>
<td>1 (100.0; 100.0)</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Non-surgical experiences</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Gaming experience (TV/video/computer)</td>
<td>9</td>
<td>673 (32)</td>
<td>7 (88.9; 95.2)</td>
<td>2 (11.1; 4.8)</td>
<td>69,46,50,51,70,60,38,37,72</td>
</tr>
<tr>
<td>14</td>
<td>Musical instrument</td>
<td>7</td>
<td>604 (0)</td>
<td>0</td>
<td>7 (100.0; 100.0)</td>
<td>71,49,50,51,70,60,38</td>
</tr>
<tr>
<td></td>
<td>Activity</td>
<td>n</td>
<td>Mean (S.D)</td>
<td>Median</td>
<td>Range (Min-Max)</td>
<td>References</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------</td>
<td>---</td>
<td>------------</td>
<td>--------</td>
<td>----------------</td>
<td>------------</td>
</tr>
<tr>
<td>15</td>
<td>Typing</td>
<td>3</td>
<td>175 (0)</td>
<td>0</td>
<td>3 (100.0; 100.0)</td>
<td>49, 50, 38</td>
</tr>
<tr>
<td>16</td>
<td>Sport</td>
<td>2</td>
<td>82 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>71, 38</td>
</tr>
<tr>
<td>17</td>
<td>Chopsticks</td>
<td>2</td>
<td>118 (67)</td>
<td>1 (50.0; 56.8)</td>
<td>1 (50.0; 43.3)</td>
<td>49, 50</td>
</tr>
<tr>
<td>18</td>
<td>Sewing</td>
<td>2</td>
<td>118 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>49, 50</td>
</tr>
<tr>
<td>19</td>
<td>Driving</td>
<td>2</td>
<td>100 (43)</td>
<td>1 (50.0; 43.0)</td>
<td>1 (50.0; 57.0)</td>
<td>51, 38</td>
</tr>
<tr>
<td>20</td>
<td>Experience operating tools</td>
<td>2</td>
<td>118 (0)</td>
<td>0</td>
<td>2 (100.0; 100.0)</td>
<td>49, 50</td>
</tr>
<tr>
<td>21</td>
<td>Billiards</td>
<td>1</td>
<td>21 (21)</td>
<td>1 (100.0; 100.0)</td>
<td>0</td>
<td>72</td>
</tr>
</tbody>
</table>

**Surgical experiences**

<table>
<thead>
<tr>
<th></th>
<th>Activity</th>
<th>n</th>
<th>Mean (S.D)</th>
<th>Median</th>
<th>Range (Min-Max)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Laparoscopic experience</td>
<td>2</td>
<td>46 (25)</td>
<td>1 (50.0; 54.3)</td>
<td>1 (50.0; 45.65)</td>
<td>71, 72</td>
</tr>
<tr>
<td>23</td>
<td>Endovascular experience</td>
<td>1</td>
<td>61 (0)</td>
<td>1 (100.0; 100.0)</td>
<td>0</td>
<td>46</td>
</tr>
</tbody>
</table>

*Studies - references within the body of the manuscript*
### Table 3: A summary of visual spatial tests as predictors of surgical performance

<table>
<thead>
<tr>
<th>Count</th>
<th>Visual spatial test (n=25)</th>
<th>Description of innate ability</th>
<th>Number of studies</th>
<th>Total number of participants (number of participants in significant studies)</th>
<th>Number of studies reporting a significant association (percent; weighted percent)</th>
<th>Number of studies that failed to identify a significant association (percent; weighted percent)</th>
<th>Number of studies stratified by type of technical skill</th>
<th>Number of studies reporting a significant association stratified by technical skill category (percent)</th>
<th>Number of studies that failed to identify a significant association stratified by technical skill category (percent)</th>
<th>*Studies n=38</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Card rotation</td>
<td>A picture of a 2-dimensional shape is set as the reference figure. Participants are then required to indicate whether they need to flip or rotate eight additional 2-dimensional shapes in order to match the reference figure. The test is completed in a set time and a score is generated on accuracy.</td>
<td>10</td>
<td>229 (68)</td>
<td>5 (50.0;30.0)</td>
<td>5 (50.0;70.0)</td>
<td>Open surgery 1 0 (0.0) 1 (50.0) 1 (100.0)</td>
<td>75,73,72,72</td>
<td>Open surgery 1 0 (0.0) 1 (50.0) 1 (100.0) 75,73,72</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Laparoscopic skills 2 3 (50.0) 3 (50.0) 1 (50.0)</td>
<td>40,77,69,64</td>
<td>Laparoscopic skills 2 3 (50.0) 3 (50.0) 1 (50.0) 40,77,69</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR laparoscopy 6 0 (0.0) 1 (50.0) 2 (66.7)</td>
<td>52,61,62,64</td>
<td>VR laparoscopy 6 0 (0.0) 1 (50.0) 2 (66.7) 52,61,62</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR endoscopy 2 1 (50.0)</td>
<td>64</td>
<td>VR endoscopy 2 1 (50.0)</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td>2</td>
<td>Mental rotation test</td>
<td>Participants are shown two three-dimensional images and asked to compare the images and state whether they are the same image or the mirror image. The test is completed in a set time and a score is generated on accuracy.</td>
<td>9</td>
<td>389 (288)</td>
<td>6 (66.7;74.0)</td>
<td>3 (33.3;26.0)</td>
<td>Open surgery 4 3 (75.0) 1 (100.0) 1 (25.0)</td>
<td>75,76,57,5,76,45,77,64</td>
<td>Open surgery 4 3 (75.0) 1 (100.0) 1 (25.0) 75,76,57,5,76,45,77,64</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR laparoscopy 3 1 (33.3) 2 (66.7)</td>
<td>67,80</td>
<td>VR laparoscopy 3 1 (33.3) 2 (66.7) 67,80</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR endoscopy 2 1 (50.0)</td>
<td></td>
<td>VR endoscopy 2 1 (50.0)</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td>3</td>
<td>PicSOR</td>
<td>Participants are required to move a spinning arrow on top of a cube, until the angle between the two objects is 90 degrees. The closer the approximate angle is to the actual angle the higher the score.</td>
<td>8</td>
<td>255 (206)</td>
<td>5 (62.5;80.8)</td>
<td>3 (37.5;19.2)</td>
<td>Laparoscopic skills 3 2 (66.7)</td>
<td>63,73,39,74,72,60,61,62</td>
<td>Laparoscopic skills 3 2 (66.7) 1 (25.5) 3 (75.0) 63,73,39,74,72,60,61,62</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR laparoscopy 4 3 (75.0) 1 (25.5)</td>
<td>61,62,64</td>
<td>VR laparoscopy 4 3 (75.0) 1 (25.5) 61,62</td>
<td>*Studies n=38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>VR endoscopy 2 0 (0.0)</td>
<td></td>
<td>VR endoscopy 2 0 (0.0)</td>
<td>*Studies n=38</td>
</tr>
</tbody>
</table>
4 Cube comparison  Participants are asked to compare two three dimensional cubes with a letter on each surface and indicate whether they are the same cube or a different cube. The test is completed in a set time and a score is generated on accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
<th>VR endoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>181 (55)</td>
<td>3 (37.5;30.4)</td>
<td>5 (62.5;69.6)</td>
<td></td>
</tr>
</tbody>
</table>

5 Map planning  Participants need to find the shortest route between two points while avoiding road blocks and passing along the side of the building. This task is scored on time and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
<th>VR endoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>163 (74)</td>
<td>4(50.0;45.4)</td>
<td>4 (50.0;54.6)</td>
<td></td>
</tr>
</tbody>
</table>

6 Surface development  Participants are given a 2-dimensional blueprint to fold in three-dimensional space and match to the appropriate 3-dimensional picture. This task is scored on time and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
<th>VR endoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>164 (91)</td>
<td>3(60.0;44.5)</td>
<td>2 (40.0;55.5)</td>
<td></td>
</tr>
</tbody>
</table>

7 Minnesota paper form board test  Participants are given a set of different parts and are required to choose which of the 5 arrangements could be made up of these parts. This task is scored on time and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
<th>VR endoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>339 (157)</td>
<td>2 (40.0;46.3)</td>
<td>3 (60.0;53.7)</td>
<td></td>
</tr>
</tbody>
</table>

8 Thurstone hidden figures test  Participants are given a reference image and are required to indicate whether the reference image is imbedded in the subsequent more complex image. Time and number of correct responses is the final score.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>344 (73)</td>
<td>1 (20.0;27.2)</td>
<td>4 (80.0;78.8)</td>
</tr>
</tbody>
</table>

9 Paper folding  A piece of paper is folded in a specific way and then hole punched through. The participant uses this reference image to match which piece of paper would correspond to the paper when unfolded. Time and number of correct responses is the final score.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>121 (48)</td>
<td>1 (33.3;39.7)</td>
<td>2 (66.7;60.0)</td>
</tr>
</tbody>
</table>

10 Gestalt completion test  Participants are given a drawing of a fragmented object and are required to try and identify what it is. Time and number of correct responses is the final score.

<table>
<thead>
<tr>
<th></th>
<th>Open surgery</th>
<th>Laparoscopic skills</th>
<th>VR laparoscopy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>191 (107)</td>
<td>1(33.3;56.0)</td>
<td>2 (66.6;44.0)</td>
</tr>
<tr>
<td></td>
<td>Test Name</td>
<td>Description</td>
<td>Number of Participants</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>11</td>
<td>Guay’s visualization of viewpoints test</td>
<td>Participants are shown a picture of a cube with an image inside the cube. They are then shown the same cube in a different orientation and required to identify the corners of the cube where the new view was taken. Time and number of correct responses is the final score.</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>Rey-Osterrieth complex figure test</td>
<td>Participants are asked to refer to a complex line drawing and first draw it out by referring to the original image and then again by memory. Accuracy is the final score.</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>Space relations test</td>
<td>Participants are required to compare letters, numbers and/or objects quickly and accurately in a certain amount of time. Time and number of correct responses is the final score.</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Perceptual speed</td>
<td>Participants are given a two-dimensional pattern and asked to fold it and rotate the shape in order to match it to a three-dimensional image. Time and number of correct responses is the final score.</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>4 &quot;magic eye&quot; images</td>
<td>Participants are given four separate three dimensional images hidden within a 2-dimensional pattern and are asked to identify each image. Time and number of correct responses is the final score.</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>Adapted Corsi Block Tapping Test</td>
<td>A monitor displays nine dice cubes in random position. A certain number of these dice are highlighted in a given order, the participant is then required to point out these dice in the same order. If more than three are correct the number of dice increases on the subsequent turn.</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>Choosing a path</td>
<td>The participant is given a diagrammed imagine of a city map. They must then plan routes between two set points</td>
<td>1</td>
</tr>
</tbody>
</table>
and avoid any roadblocks. This task is scored on time and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Task Description</th>
<th>Participants</th>
<th>Test Environment Options</th>
<th>Score 1</th>
<th>Score 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Five porteus maze</td>
<td>1</td>
<td>140 (0)</td>
<td>0</td>
<td>1 (100.0;100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Open surgery</td>
<td>2</td>
<td>0 (0.0)</td>
<td>2 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>55.75</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Matrix reasoning</td>
<td>1</td>
<td>21 (0)</td>
<td>0</td>
<td>1 (100.0;100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laparoscopic skills</td>
<td>1</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VR laparoscopy</td>
<td>1</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>57</td>
<td>72</td>
</tr>
<tr>
<td>20</td>
<td>Phase discrimination test</td>
<td>1</td>
<td>47 (0)</td>
<td>0</td>
<td>1 (100.0;100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Open surgery</td>
<td>1</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Shape memory test</td>
<td>1</td>
<td>37 (0)</td>
<td>0</td>
<td>1 (100.0;100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Open surgery</td>
<td>1</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Touching blocks test</td>
<td>1</td>
<td>107 (107)</td>
<td>1 (100.0;100.0)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laparoscopic skills</td>
<td>1</td>
<td>1 (100.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Tube shape test</td>
<td>1</td>
<td>58 (0)</td>
<td>0</td>
<td>1 (100.0;100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>VR endoscopy</td>
<td>1</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82</td>
<td></td>
</tr>
</tbody>
</table>
Participants are asked to identify a picture covered with a visual obstruction, as quickly as possible. Time and number of correct responses is the final score.

A picture of a side view, top view and bottom view are given to the participant who must then visualize these views in three dimensions to carve out a cylinder-shaped piece of soap. A score is given for the accuracy of the final soap carving.

**Table 4: Summary of dexterity tests as predictors of surgical performance**

<table>
<thead>
<tr>
<th>Count</th>
<th>Dexterity test</th>
<th>Description of innate ability</th>
<th>Total number of participants (number of participants in significant studies)</th>
<th>Number of studies reporting a significant association (percent; weighted percent)</th>
<th>Number of studies that failed to identify a significant association (percent; weighted percent)</th>
<th>Number of studies stratified by type of technical skill</th>
<th>Number of studies reporting a significant association stratified by technical skill category (percent)</th>
<th>Number of studies that failed to identify a significant association stratified by technical skill category (percent)</th>
<th>*Studies n=19</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Purdue pegboard</td>
<td>This task is completed with a board with two parallel rows of holes (25 in total). Using their hands, the participant is required to place cylindrical shaped pegs into the holes on the board as quickly as possible.</td>
<td>6  399 (56)</td>
<td>2 (33.3;14.0)</td>
<td>4 (66.7;86.0)</td>
<td>Open surgery 4  1 (25.0)</td>
<td>3 (75.0)</td>
<td>3,65,66,72,5</td>
<td>1 (100.0)</td>
</tr>
<tr>
<td>2</td>
<td>Grooved peg board</td>
<td>A square board with 25 holes (5 holes across and 5 holes down) is used. Using their hands, the participants are required to place the metal pegs into the holes on the board as quickly as possible.</td>
<td>5  95 (74)</td>
<td>4 (60.0;77.9)</td>
<td>1 (40.0;22.1)</td>
<td>Laparoscopic 1  0 (0.0)</td>
<td>1 (100.0)</td>
<td>0 (0.0)</td>
<td>5,61</td>
</tr>
</tbody>
</table>
3 Crawford Small Parts Dexterity test

A board with holes for pins and screws is presented to the participants. Using tweezers, the pins are placed into their designated holes and collars are placed on the pins sticking out. The screws are then placed into their designated holes and a screwdriver is used to screw down each screw until flush with the board. Participants are asked to complete the task as quickly as possible.

4 ADTRACK2

A screen displays a block bounded by two bars. The bars move from side to side and the participant uses a joystick to keep the block as close to the bars as possible. The closer the block stays to the bars, the higher the score.

5 Block Design

Participants are given a pattern and are required to rearrange blocks with their hands in order to recreate the pattern. The final score is calculated from accuracy and speed.

6 Porteus Maze

Participants are required to trace through a maze from one end to the other, avoiding dead ends or backtracking. A time limit is set depending on the complexity of the maze.

7 Tactual Performance

Participants are blindfolded and asked to place a number of cut out shapes into their corresponding positions on a form board. The participant completes the task three times; with their left hand, right hand and both hands together. Time is used for the final score.

8 ADEPT

Participants read off a computer screen while working within a dome with laparoscopic instruments and a camera. Four tasks are completed. The error plate detects excessive movement and a score is generated.

9 ADTRACK3

A screen displays a block bounded by two bars. The bars move from side to side and the participant uses a joystick to keep the block as close to the bars as possible. The task differs from ADTRACK2 in that the participant can adjust the difficulty. The higher the level and the closer the block stays to the bars, the higher the score.
<p>|   | Test Name                                      | Description                                                                                                                                                                                                 | Score | Time | Score Type       | Score | Time | Score Type       | Score | Time | Score Type       | Score | Time | Score Type       |
|---|-----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|-------|------|------------------|
|10 | Bimanual coordination test                    | The participant uses two controls to navigate two points through a labyrinth. The right hand moves the points up and down and the left moves the points side to side. Accuracy and time generate a final score. | 58 (0) | 0    | 1 (100.0;100.0)  | VR endoscopy | 1    | 0 (0.0)          | 1 (100.0) | 82  |
|11 | Bennett hand tool dexterity test              | Screws, nuts and bolts are secured tightly into a wooden frame. Using the tools provided the participant is required to take apart the units and reassemble them on the opposite side. Time is used for the final score. | 8 (0)  | 0    | 1 (100.0;100.0)  | Open surgery | 2    | 0 (0.0)          | 2 (100.0) | 66  |
|12 | Reaction time                                 | Participants hold down a button; then when a neighboring button lights up they are required to tap it. Cumulative response time is used for the score.                                                              | 21 (0) | 0    | 1 (100.0;100.0)  | VR laparoscopy | 1    | 1 (100.0) | 0 (0.0) | 72  |
|13 | Double labyrinth test                         | The participant uses two levers to control the position of two markings within a cylinder, which rotates at a constant speed. The right hand controls the right point and the left hand controls the left point. If a marking touches the side of the screen then an error is recorded. | 58 (58) | 1 (100.0;100.0) | 0   | VR endoscopy | 1 (100.0) | 0 (0.0) | 82  |
|14 | Finger tap test                               | Participants place their hand palm down over a board with their index finger overlaying a device that counts the number of finger taps. For 10 seconds at a time they are required to tap as quickly as possible for 3-6 repetitions. This test is completed with the left and right hand. The more finger taps completed in the allotted time the higher the score. | 12 (21) | 0    | 1 (100.0)        | Laparoscopic skills | 1    | 0 (0.0) | 1 (100.0) | 72  |
|15 | Gibson spiral maze test                       | Participants are given a circular maze and are required to trace between the lines. Score is determined by time and error.                                                                                  | 10 (0) | 0    | 1 (100.0;100.0)  | Open surgery | 1    | 0 (0.0) | 1 (100.0) | 36  |
|16 | Minnesota rate manipulation test              | The participant is required to place a number of discs onto a large board in a specific series. Once completed, the participant then flips over each disc in series moving along each row in a consecutive manner until all the discs have been flipped. | 8 (0)  | 0    | 1 (100.0;100.0)  | Open surgery | 1    | 0 (0.0) | 1 (100.0) | 66  |
|17 | Steadiness hole test                          | The participant is required to place a metal tip stylus into 9 progressively smaller holes without touching the edges. This task is scored on time and error.                                                | 8 (8)  | 1 (100.0;100.0) | 0 | Open surgery | 1 | 0 (0.0) | 1 (100.0) | 66  |</p>
<table>
<thead>
<tr>
<th>Test</th>
<th>Description</th>
<th>Score</th>
<th>Time</th>
<th>Notes</th>
<th>Laparoscopic Skills</th>
<th>VR Laparoscopy</th>
<th>ADTRACK</th>
<th>ADTRACK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tremor</td>
<td>Participants grab a needle with a laparoscopic instrument that is attached to an oscillator and hold it steady for 20 seconds. Steadiness is used as the final score.</td>
<td>1</td>
<td>21 (0)</td>
<td>0 (0.0)</td>
<td>1 (100.0;100.0)</td>
<td>1 (0.0)</td>
<td>1 (100.0)</td>
<td></td>
</tr>
<tr>
<td>Wire Loop Dexterity Test</td>
<td>The participant is required to pass a hand held loop over a wire that has three formed bends. The goal of the task is to move from one end of the wire to the other without making contact with the loop.</td>
<td>1</td>
<td>37 (37)</td>
<td>0 (0.0)</td>
<td>Open surgery</td>
<td>0 (0.0)</td>
<td>1 (100.0)</td>
<td></td>
</tr>
</tbody>
</table>

ADEPT - Advanced Dundee Endoscopic Psychomotor Tester; VR – Virtual Reality; ADTRACK - Adaptive Tracking Task; *Studies - references within the body of the manuscript
2.3.2.4 Discussion

Training residents requires considerable trainee and faculty commitment and substantial fiscal resources. Furthermore, selection for technical aptitude has gained increased relevance amongst other selection criteria given resident work hour restrictions, growing complexity of surgical procedures and increasing incorporation of more difficult minimally invasive techniques into daily practice (Biondi et al., 2013; Khatuja et al., 2014; Levine & Spang, 2014). Therefore, if predicting technical aptitude were possible, this component could be added to the current selection process to benefit all parties involved. It is therefore not surprising that a large amount of effort has been dedicated to identifying an objective test to measure the underlying innate abilities thought to be required for trainees to reach competent technical performance. Despite 52 studies that have investigated 23 participant characteristics, 25 visual spatial tests, 20 dexterity tests, 3 basic performance resource tests and 12 other innate ability tests, no single test or combination of tests has been identified to validly and reliably predict technical aptitude.

Of the surrogate markers investigated by participant questionnaires and dexterity testing, only a small minority regularly demonstrated a positive relationship with technical performance. Notable among these are previous video game playing experience, and performance on the grooved pegboard dexterity test. Although some authors have considered video game experience as a potential marker of superior innate technical ability, evidence suggests that video game playing experience may be more appropriately considered a learned skill that only confers a short-term advantage. Gamers appear to demonstrate superior technical ability, and this may be the result of their familiarity with reacting to 2-dimensional feedback from a monitor using unwatched hands, as well as increased attention and faster reaction time (Green & Bavelier, 2003; Hislop et al., 2006; Rosser et al., 2007). However, Van Hove et al. reported that the gaming advantage was lost after initial baseline testing, and Dimitriou et al. further noted that gamers’ technical skill improved slower when compared to non-gamers (Dimitriou et al., 2009; Van Hove et al., 2008). Together, these findings suggest that the initial gaming performance advantage can be quickly overcome by minimal technical skills training. Similarly, technical skills training may also override any baseline advantage associated with higher performance on manual dexterity testing. Van Herzeele et al. reported a significant correlation between manual dexterity as quantified using the grooved pegboard test, and baseline performance on a virtual
reality endovascular renal artery stenosis task. However, the predictive value of the dexterity text was lost after participants were given the opportunity to practice this task (Masud et al., 2012). Thus, it appears that some surrogate tests that have been presumed to reflect innate ability may in fact be assessing learned skills, limiting their value as predictors of technical performance by surgical trainees.

It would seem intuitive that tests of visuospatial ability would reliably predict technical ability, given surgeons’ need to operate around complex structures in three-dimensional space. One might expect this relationship to be particularly strong in the case of laparoscopic procedures, where surgeons operate in three dimensions while relying on two-dimensional visual feedback. Therefore it is particularly surprising that of the 26 visual spatial tests studied to date, only 3 (12%) (the Card Rotation Test, the Mental Rotation Test (MRT), the PicSOR test) have consistently demonstrated a positive relationship with technical ability.

While there may be a perception that visuospatial ability is a single homogenous aptitude that varies between individuals, the results of the present review suggest that different visuospatial tests may assess distinct and discrete abilities, each of which may only apply to a specific subset of surgical procedures. For example, while the MRT was positively correlated with technical performance in the majority of studies, this association differed depending on the type of surgical task. MRT results significantly predicted technical performance in all studies involving open surgical tasks in a simulated environment. A positive correlation was seen between MRT scores and performance of internal fixation of a mandible fracture (Wanzel et al., 2003), 4 flap z-plasty, (Wanzel et al., 2002), and with tying of an open surgical reef knot (Brandt & Davies, 2006). In contrast, when assessed as potential predictors of simulated endoscopic and laparoscopic tasks performance, the results were inconsistent, with half of studies finding no predictive value (Groenier et al., 2014; Hedman et al., 2006; M. M. Keehner et al., 2004; Luursema et al., 2010). Similar discrepancies are noted with the predictive value of the PicSOR test to predict minimally invasive performance. Although this test was developed specifically to test the visual spatial abilities required for laparoscopy, one might expect it to also be equally effective in predicting technical performance on endoscopic tasks, given the similar minimally invasive nature of both. However, PicSOR has been successful in predicting simulated laparoscopic tasks but not simulated endoluminal tasks in studies to date. Specifically, PicSOR has demonstrated a positive correlation with a laparoscopic cutting task (A. G. Gallagher et al., 2003), manipulation and
diathermy task (McClusky et al., 2005), and laparoscopic peg transfer in the novice trainee (Kolozsvari et al., 2011), but no correlation with the endobubble and gastroscopy tasks on the virtual reality simulator (L. Enochsson et al., 2006; E. Matt Ritter et al., 2006). Therefore, although visual spatial ability is likely important for technical performance, the surrogate tests available seem to predict performance on specific subgroups of surgical tasks instead of reflecting underlying abilities that transfer across surgical procedures in general.

Given that the purpose of this systematic review was to focus on aptitude testing for selection into training, open and minimally invasive techniques have been discussed together. This is appropriate for contemporary surgical practice, where a number of specialties (e.g. general surgery, urology) require that surgeons be proficient with a mix of both open and minimally invasive techniques, and this is also reflected in training curricula. However, the literature does suggest that different tests may be more appropriate for predicting aptitude for either open or minimally invasive techniques. These tests may ultimately be able to provide some insight into whether surgeons would benefit from biasing their independent practices following completion of training toward either open or minimally invasive procedures. Furthermore, some surrogate markers seem to be more appropriately suited for specific surgical specialties. For example MRT has been shown to be predictive for z-plasty, which is pertinent for reconstructive surgery. However, this same correlation has not been demonstrated for open abdominal surgery (Deary et al., 1992; Schueneman et al., 1984). Similarly, PicSOR has demonstrated some predictive potential for laparoscopic tasks and therefore may be more useful in surgical disciplines where laparoscopic techniques are common (e.g. urology, general surgery or obstetrics and gynecology). It is therefore, possible that a different set of surrogate tests may be required for selection into different surgical specialties.

While several studies have examined the association between surrogate marker scores and cross-section performance on one or more surgical tasks, there is limited evidence concerning potential associations between natural aptitude and longitudinal performance in terms of either rate of skills acquisition or performance over the longer term. Buckley et al. demonstrated that after testing 86 novices, 12 high aptitude and 12 low aptitude scoring individuals correlated to fast versus slow learners when completing simulated general surgery tasks. This suggests that aptitude tests may be able to select out the gifted and the slow learner, however the surgical performance of individuals at these extremes was not studied. In addition, no study has evaluated
whether these initial predictions in the simulation setting transfer to the real operating environment. Therefore, further longitudinal study is needed to better delineate the value of aptitude testing in terms of predicting both initial performance level as well as longer-term skills acquisition and performance in the real operating room, before this testing can be used to select candidates entering and progressing through surgical training.

Another potential benefit of aptitude testing might be to objectively identify trainees’ technical strengths and weaknesses, allowing for the tailoring of training activities. However, to the best of our knowledge, no studies to date have specifically evaluated this role for aptitude testing. Furthermore, it is unclear whether the use of surrogate markers to identify trainees’ strengths and weaknesses confers any benefit over the use of existing in-training evaluation tools (e.g. in-training evaluation reports, objective structured assessments of technical skills, procedure-specific checklists etc.). Thus, further study is required to determine whether these tests can be used in this way, including optimal testing frequency (once versus several occasions), the need to control for time in training to account for potential change in aptitude scores over time, and benefits of surrogate marker testing versus in-training tools.

Ultimately, technical performance is likely a result of a complex interplay of numerous innate abilities, in conjunction with subsequent exposures and experiences. These may vary between individuals, potentially explaining the limited success to date in identifying individual surrogate markers that reliably predict technical performance. In an attempt to account for these complexities, other performance disciplines utilize a battery of tests in an attempt to simultaneously assess a range of innate abilities, with the goal of selecting candidates based on the cross-section of innate abilities believed to maximize performance in a given field. For example, the United States military uses a testing battery, called the Armed Services Vocational Aptitude Battery, to characterize new recruits across a broad range of innate abilities. (Mayberry & Carey, 1997) This test battery evaluates arithmetic reasoning, word and mathematics knowledge and paragraph comprehension. The results are then used to direct recruits toward specific branches and roles within the military (e.g. pilot training, intelligence) that best reflect their set of innate abilities. Perhaps motivated by this experience, Deary et al. (Deary et al., 1992) and Schueneman et al. (Schueneman et al., 1985; Schueneman et al., 1984) developed and evaluated test batteries for selection into surgery. They tested combinations of characteristics including dexterity, visual spatial ability, and personality traits, but failed to find a combination
tests that predicted surgical performance in the operating room as assessed by staff surgeons. More recently, Gettman et al., (Gettman et al., 2003) Johnson et al., (Johnson et al., 2004) and Matsumoto et al. (Matsumoto et al., 2006) have tried a more sophisticated approach to predict performance across a range of surgical procedures by combining the assessment of multiple abilities into a single model, known as Basic Performance Resources (BPR). This model involves measuring a range of innate abilities such as visual-hand response speed, visual information processing speed and grip strength. The results are then analyzed using nonlinear causal resource analysis that attempts to identify the point at which an individual’s performance resources become insufficient to meet those required by a given surgical task. Resources become insufficient when one or more of the above abilities is unable to meet the challenge of the surgical performance. In the three studies of BPR reported to date, this method has accurately predicted technical performance for 62 to 75% of participants (Cadeddu & Kondraske, 2007; Gettman et al., 2003; Johnson et al., 2004; Matsumoto et al., 2006). Although not perfect, this model is the first to attempt to quality the interplay of many abilities that likely work together to produce an individuals end performance. BPRs also acknowledge that the reasons for weak performance are likely due to a different underlying ability, depending on the performer. Therefore, by testing multiple abilities the BPR approach may have a higher likelihood of understanding where restrictions most often occur. However, given the limited number of studies and the small sample sizes used, further BPR studies are required to optimize the combination of innate abilities best suited for predicting technical performance for different surgical techniques to increase the accuracy of predictions. To date, this method has not been used in open surgery and may be worth pursuing. Furthermore, further study is required to confirm its value and feasibility as a selection tool for surgical training.

Given the large number of surrogate markers evaluated to date, it may be beneficial for future studies to focus on the surrogate tests that have reported positive associations. Based on the studies identified through in this review, the majority have reported positive associations between PicSOR and laparoscopic procedures. Thus it would appear to be reasonable to include the PicSOR in studies evaluating associations with performance on laparoscopic tasks. Similarly, the MRT has shown positive associations with z-plasty and mandibular bone fixations, therefore, surgical procedures involving other bone fixation or flap reconstruction maneuvers may be more likely to correlate with this specific visual spatial test. However, in our opinion, because of a lack
of any other consistently reported associations, there is insufficient evidence to recommend any other test for studies that involve any other open or minimally invasive surgical task. With respect to combination studies, future work may benefit from incorporating these potentially promising individual surrogate tests, as part of a combination that is appropriate for a specific surgical discipline. Of the combination studies identified in this systematic review, Deary et al. used the MRT but did not find a significant correlation with performance in the operating room. However, this work was done before the development of surgical simulation and a limitation of this study was the use of subjective assessment tools. PicSOR was developed after both these studies had been published and therefore was not used in either one. Given the advances in surgical education with respect to simulation and objective performance assessment tools, future combination studies may confer more promising results when incorporating these advances. Alternatively, BPR offers the novel advantage of attempting to quantify the interplay of many abilities to predict each individual’s end performance, and more work with this model may prove beneficial.

A number of studies have investigated the relationships between personal characteristics and/or cognitive test results, and surgical performance. However, to date, no single test has been reported to reliably predict technical performance across the range of techniques and skills required of surgical trainees. Visual spatial tests have demonstrated some promise, but only in predicting performance on a specific subset of surgical tasks. It appears that strategies such as BPRs, that assess multiple innate abilities, their interaction, and their relationship with technical skill, may be more likely to ultimately serve as reliable predictors of future surgical performance. However, studies of this nature are limited to date, and therefore more robust research is required before implementing this method into the surgical trainee selection processes.

2.3.3 The role of simulated technical tasks to inform trainee selection

An alternative approach to surrogate markers in predicting technical aptitude may be to incorporate simulated technical tasks into the selection process. The United Kingdom and Ireland both incorporate simulated technical tasks into their selection process for advanced surgical training. The surgical training system in these countries differs from that of North America and
instead of medical students directly entering 5 years of surgical residency, their programs are separated into basic and advanced surgical training. After medical school students interested in a surgical career enter basic surgical training, which has a focus in surgery but is a rotating internship that teaches the foundation of all clinical practices and lasts two years (Beard, 2008; A. G. Gallagher et al., 2008). At the completion of these two years students then apply for specialty surgery training, which involves a written application, knowledge examination and technical skills bell ringer exam (Evgeniou, Peter, Tsironi, & Iyer, 2013). If they are not successful in entering specialty surgical training, the student then enters a non-surgical career with no consequences.

The technical examination in these countries is a series of simulated technical tasks assessed by the surgical faculty for the open tasks, and the virtual reality simulator for the laparoscopic tasks. Examples of open tasks included in the technical skills exam in Ireland are simulated end-to-end bowel anastomosis, or resection of an ingrown toenail. In contrast, examples of minimally invasive skills tested on the virtual reality simulator are upper GI endoscopy, or core laparoscopic skills (A. G. Gallagher et al., 2008). The skills exam was designed to ensure that students have the technical aptitude required before entering specialty training. Unfortunately, no long-term follow up data has been published quantifying whether this objective skills test predicts clinical performance during training. However, given that many simulation studies have demonstrated transfer of skills into the real operating room (Palter & Grantcharov, 2012; Zevin B, Dedy NJ, Bonrath EM, & TP., 2013) and that the simulation tasks used in this test are directly linked to skills performed in the real clinical setting, it may be reasonable that scoring well on these skills may transfer to the clinical setting.

The inherent structure of the North American selection process does not allow for a two-tiered scheme with a simulation exam after two year. However, incorporating technical skills during the selection process may still be possible. A single program in the United States has incorporated a technical task into their selection process at the time of the in-person interview. On interview day, the Otolaryngology program at the Mayo Clinic School of Medicine, has their applicants complete a simple suture microsurgical task. The students are oriented to the microscope, instrument handling and basic microvascular knot tying technique. The students are given 20 minutes to close a vertical incision in a nylon glove, with simple interrupted knots using 10-nylon suture. A plastic surgeon on faculty grades performance in real time using a global
rating scale (Carlson, Archibald, Sorom, & Moore, 2010). Longitudinal follow up of performance scores on this screening tool and in-hospital clinical performance suggest that there is a moderate association between the global rating scale (GRS) scores of the residents who successfully entered the program and faculty clinical performance score after graduation (Moore, Price, Van Abel, & Carlson, 2014). Unfortunately, there is no data comparing students who performed poorly on the screening assessment and their clinical performance during residency.

However, of two methods of incorporating technical aptitude into selection, the UK and Irish program allow students to learn technical skills over time, which is more representative of how learning occurs in the true training environment. The disadvantage of the Otolaryngology program is that they assess the applicant for a single moment in time on interview day which may not be representative of overall performance. To incorporate the acquisition of technical skills over time into the Canadian system, assessing technical skill progression at the medical student level may be most feasible. Therefore, a simulation curriculum to screen for technical aptitude was designed for medical students. The components, structure and evidence supporting the curriculum design are explained below.

### 2.4 Simulation modalities for technical skill training

Studying the learning curves of medical students performing simulated technical tasks learned within a standardized curriculum may provide evidence to identify students who will excel or have difficulty with technical skills during residency. However, before designing a simulation curriculum and introducing a simulated technical task to a trainee it is important to understand the full breadth of simulation evaluable and which skills are most suitable for simulation training. The following elements should be considered: 1) the type of simulation model and its appropriateness for the technical task, 2) the cost of the model, and 3) the experience level of the trainee (e.g. junior vs. senior trainee).
2.4.1 Open simulation models

A number of different simulation models have been reported for the training of open surgical tasks (i.e. those performed through a traditional surgical incision providing direct visualization of the operative field). These can be stratified into man-made or non-living tissue models (low or high fidelity models), and live tissue models (high fidelity models).

2.4.1.1 Low and high fidelity simulation models

Surgical simulators designed for open surgical tasks can be separated into low fidelity and high fidelity synthetic models and high fidelity cadaveric tissue or live tissue models. Low fidelity synthetic models are generally made from plastic, paper or felt and are typically used to teach basic technical skills such as knot tying or simple suturing techniques (Hammoud et al., 2008). These models are generally reusable and are lower in cost (Figure 2a). Low fidelity models have been shown to be equivalent to high fidelity models when teaching novice trainees (Grober et al., 2004). Furthermore low fidelity models have been shown to improve technical performance within the operating room, and therefore are an appropriate and effective choice for teaching medical students and junior trainees (Chipman & Schmitz, 2009; Jensen et al., 2008). In Chapter 5 ‘low-fidelity models were chosen for the open tasks given their cost advantage and equivalence to high fidelity models when used by junior trainees, as described in section 2.5.1’

High fidelity models, however, have been shown to be superior for higher-level trainees (e.g. senior resident) who have more experience working with real tissue (Shetty, Zevin, Grantcharov, Roberts, & Duffy, 2014; Sidhu, Park, Brydges, MacRae, & Dubrowski, 2007). Common high fidelity models include cadaveric porcine and canine models and intricate synthetic models (Figure 2 b,c). These models are often not reusable and costly for ongoing training curricula.
2.4.1.2 Live models

Live models such as pigs or dogs are the most realistic simulation models available, and afford the opportunity to mimic surgical complications such as bleeding. However, they are costly and require resources such as a licensed animal laboratory, specially trained personnel to care for the animals, and appropriate anesthesia (Baddoo et al., 2008). Operative equipment and surgeon teachers’ time are also costly. Therefore, live models are typically reserved for circumstances where other simulation models are not adequate. Two examples where simulation may not be sufficient include: 1) the Advanced Trauma and Operative Management (ATOM) course, which allows senior level surgical trainees to learn rare but life-saving operative techniques such as the management and repair of stab wounds to the chest and heart, and 2) the Society of American Gastrointestinal and Endoscopic Surgeons resident courses, designed to teach advanced minimally invasive procedures that are very difficult to simulate (Ali, Ahmed, Jacobs, & Luk, 2008; S. o. A. G. a. E. Surgeons, 2015).
2.4.2 Minimally-invasive simulation models

Simulation models for minimally-invasive surgical tasks (ie. those relying on indirect visualization of the operative field through a laparoscope, endoscope, or other similar device) can be stratified into those that rely on visualization of an existing three dimensional simulated operative field, and those that use a computer-generated virtual space.

2.4.2.1 Laparoscopic box trainers

Box trainers (BT) are used to simulate basic and advanced laparoscopic tasks using either low or high fidelity models (Figure 3). The BT is a plastic box with multiple openings to allow the introduction of the simulation model, and port site openings to insert the laparoscopic camera that projects onto a monitor. The surgeon inserts real laparoscopic instruments into the box to complete the task, and operates while viewing 2-dimensional feedback from the monitor (Palter & Grantcharov, 2010).

The most widely recognized formalized training curriculum that uses the BT is the Fundamentals of Laparoscopic Surgery (FLS) curriculum. FLS is designed to teach basic laparoscopic skills using the BT to simulate 5 tasks: peg transfer, pattern cut, placing a ligation loop, intracorporeal knot tie and extracorporeal knot tie (Fried et al., 2004). The skills learned from FLS have been shown to transfer to the real operating room (McCluney et al., 2007; Stelzer, Abdel, Sloan, & Gould, 2009).

Participation in FLS is among the graduation requirements for general surgery trainees mandated by the United States Accreditation Council for Graduate Medical Education (ACGME) (Brown, Helmer, Yates, & Osland, 2012). Although this course is also recommended by a number of general surgery training programs in Canada, it remains optional at the present time.
Figure 3: (a) the laparoscopic box trainer can be used for both basic and advanced laparoscopic tasks, using either a low fidelity model such as (b) a plastic penrose drain, or a high fidelity model such as (c) cadaveric porcine small bowel with attached mesentery.

2.4.2.2 Virtual reality simulation

Virtual reality simulation for surgical technical skill is akin to playing a video game, and is considered a high fidelity model. Virtual reality has a number of advantages and disadvantages as an educational tool (Chou & Handa, 2006). An advantage of the VR simulator is that it allows for independent study appropriate to the student’s training level. The student is able to select a task (e.g. lifting and grasping) or procedure (e.g. laparoscopic cholecystectomy) best suited for
his/her training level from a series of modules (Figure 4). The simulator can be programmed to provide instructions for the task, and generates immediate feedback scores and graphs for the student to monitor his/her progress. This interactive interface decreases the number of live instructors required to teach these skills, while providing a low stakes, stress free environment in which to learn these skills. From a research perspective, a major advantage of this system is that no personnel are required to watch and rate the procedures.

However, there are two major disadvantages to the VR: 1) a lack of realism and 2) high cost. Although the VR simulation is categorized as a high fidelity model and is designed to imitate the real surgical environment (e.g. abdomen or pelvis), the appearance is often not realistic. Furthermore, the surgeon’s interaction with the simulated tissue, the way it moves or bleeds, and the lack of tactile feedback also negatively affect the experience (Satava, 2001). Newer models have been designed to improve realism by incorporating haptic feedback. However, despite these advances, the technology has yet to meet real operating room standards. Additionally, the cost of a VR simulator is substantial, with the initial acquisition cost ranging from 77,500 to 150,000 Canadian dollars, with further requirements for ongoing support, upgrades and maintenance to keep the technology updated and running smoothly (Lowry, Porco, & Naseri, 2013; Orzech, Palter, Reznick, Aggarwal, & Grantcharov, 2012).
Figure 4: Virtual reality simulator for laparoscopic technical skill training and assessment.

2.5 Instruments for the assessment of technical skill

Many assessment tools have been developed to assess technical performance for the simulation models described above, and for procedures performed in the real operating room. The major categories of assessment tools, including task specific checklists, global rating scales, task metrics and learning curve assessment, are outlined below. These tools can be used alone, or in combination to optimize the assessment goals of different technical skills curricula.
Table 5: Task specific checklist applicable to both the placement of one interrupted suture and intracorporeal knot tying

<table>
<thead>
<tr>
<th>Instrument/tie/checklist</th>
<th>Not done</th>
<th>correctly</th>
<th>Completed</th>
<th>correctly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct‘handling’ of needle ‘driver’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct‘handling’ of forceps</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loading the ‘needle’ driver at the ‘tips’ of the ‘needle’ driver</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loading the ‘needle’ at 90° degrees to the ‘needle’ driver</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loading the ‘needle’ 2/3 and 1/3</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Holding the ‘tissue’ edge with the ‘forceps’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entering the ‘tissue’ at 90° degrees</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coming through both ‘black’ dots</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Following the ‘curve’ of the ‘needle’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exiting the ‘tissue’ using the ‘needle’ driver</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exiting the ‘tissue’ by following the ‘curve’ of the ‘needle’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulling the ‘suture’ through’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leaving a ‘short’ tail</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Placing the ‘needle’ driver between ‘short’ and ‘long’ suture’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st double throw</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Square ‘knots’</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2nd single throw</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd single throw</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hold both sutures with ‘needle’ driver</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cut with ‘suture’ scissors</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Max‘Total’20:
2.5.1 Task specific checklists

These checklists are created for specific open or minimally invasive tasks and are generally formatted as a binary scale documenting whether a step of the procedure is omitted or completed. Checklists are best suited to evaluate novice or junior trainees who are focused on learning the steps of the procedure. These tools are easy to implement because the examiner requires minimal training; the binary scale is relatively intuitive, and minimal surgical judgment is needed to score performance (Regehr, MacRae, Reznick, & Szalay, 1998). Checklists can be used to provide objective feedback both in the simulation setting and in the real operating room. However, a disadvantage to checklists is that they do not measure performance nuances or technical quality, and therefore are poorly suited to more advanced trainees (Hodges, Regehr, McNaughton, Tiberius, & Hanson, 1999). Furthermore, checklists are procedure specific, and therefore cannot be re-used for different operations (Table 5).

2.5.2 Global rating scales

GRS are used to provide an overall assessment of performance. They require the examiner observe a trainee perform a procedure and uses the judgment of these individuals to rate elements of their technical skill on a Likert scale in accordance with specific descriptive anchors. Since the tool relies on the examiner’s judgment, a subjective variable is inherently incorporated into the assessment. Therefore, GRS are more reliable in the hands of experienced examiners or alternatively, trained personal.

GRS have been shown to differentiate levels of experience and can therefore be used for both novice and more advanced trainees (Regehr et al., 1998). In addition, because these scales are not procedure specific they can be used for a multitude of tasks or operations in different surgical fields as well as within the simulation setting or the real operating room. The disadvantage of GRS is that they require trained personal to watch the procedures either in real time or on video to generate a score. Both are time consuming and resource heavy, which has been reported to severely limit uptake. Table 6 is the Objective Structure of Technical skills (OSATS), which is considered the gold standard GRS to assess technical performance (Martin J.A et al., 1997).
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect for tissue</td>
<td>Frequently used unnecessary force of tissue or caused damage by inappropriate use of instruments</td>
<td>Careful handing of tissue but occasionally caused inadvertent damage</td>
<td>Consistently handled tissue appropriately with minimal damage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time and motion</td>
<td>Many unnecessary moves</td>
<td>Efficient time/motion but occasionally causes inadvertent damage</td>
<td>Economy of movements and maximum efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument handling</td>
<td>Repeatedly makes awkward moves with instruments</td>
<td>Competent use of instruments although occasionally appears stiff or awkward</td>
<td>Fluid moves with instruments and no awkwardness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of Instruments</td>
<td>Frequently asked for the wrong instrument or used an inappropriate instrument</td>
<td>Knew the names of most instruments and used appropriate instrument for the task</td>
<td>Obviously familiar with the instruments and their names</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of assistant</td>
<td>Consistently placed assistant poorly or failed to use assistant</td>
<td>Good use of assistant most of the time</td>
<td>Strategically used assistant to the best advantage at all times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow or operation and forward planning</td>
<td>Frequently stopped operating or needed to discuss next move</td>
<td>Demonstrated ability for forward planning with steady progression of operative procedure</td>
<td>Obviously planned course of operation with effortless flow from one more to the next</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of specific procedure</td>
<td>Deficient knowledge. Needed specific instruction at most operative steps</td>
<td>Knew all important aspects of the operation</td>
<td>Demonstrated familiarity with all aspects of the operation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.5.3 Task metrics

Task metrics are the most common method to evaluate technical performance because the data are objective and easy to collect (Schmitz et al., 2014). Examples of these metrics include time, instrument motion tracking and error scores (Figure 5) (A. G. Gallagher, Richie, McClure, & McGuigan, 2001). Task metrics are used in the simulated setting with the virtual reality (VR) simulator and laparoscopic box trainer (BT). The VR simulator assessment metrics have been shown to be able to discriminate performance between experts, intermediate and novice trainees. The BT tasks measure the time of the procedure or task in real-time and quality of the end product in terms of an error score which are quickly determined and successfully discriminate trainee level (A. G. Gallagher & Satava, 2002; van Dongen, Tournoij, van der Zee, Schijven, & Broeders, 2007). However, these metrics do not necessarily capture quality. For example a trainee may complete a task quickly and accurately resulting in a high score; but may also be rough, have poor respect for tissue and have suboptimal technique, none of which are measured in this scoring system. However, the evidence does suggest that when separating the metrics, time is more accurate than motion tracking when assessing performance, and that summing these metrics into an overall score is a better assessment of technical performance (D. Stefanidis, Scott, & Korndorffer, 2009).

Penalty Scores

Secure knot = 0
Slipping knot = 10
Knot comes apart = 20

Time to complete = _______ seconds

Penalty = _____ ( + ) mm. from edge of pre-drawn dots
____ ( + ) mm. gap in incision
____ ( + ) security of knot

Figure 5: Task metrics for the laparoscopic knot tie task, using the laparoscopic box trainer
2.5.4 Using learning curves to assess technical skill progression

Understanding the progression of technical skill acquisition by analyzing learning curves is an important topic in surgical education. It is becoming increasingly apparent that although all trainees may improve with dedicated practice, a subset of trainees may not meet the safety threshold (Cushchieri, 2003), defined as competence (Szasz et al., 2014). Learning curves are defined as an improvement in performance over time. Graphically, learning curves are described by three main features: 1. the initial or start point; 2. the rate of learning; and 3. expert level, where the surgeons’ learning curve stabilizes (Cook, Ramsay, & Fayers, 2004). However, progression along the learning curve continuum varies from individual to individual.

Several graphical representations have been proposed to represent the performance over time associated with successful learning (Yelle, 1979). The most popular resembles an s-shape (a “sigmoidal” curve), with an initial familiarization phase characterized by slow improvement, followed by a period of rapid improvement with steep slope, and finally a consolidation phase where performance plateaus. The familiarization phase is often short or absent for a basic task, resulting in a curve that begins with the steep slope of the rapid improvement phase (Figure 6a). However, as the task being learned becomes more complex for the individual, it demands a more substantial familiarization phase, and the full s-shape is more clearly demonstrated (Figure 6b) (Yelle, 1979). It is important to note that all individuals will experience performance variation between attempts, and thus the individual learning curve is characterized by a series of up and down fluctuations, reflecting the underlying s-shaped distribution of performance combined with the random inter-attempt variance (Figure 6c).
Figure 6: Examples of expected smoothed learning curves for the performance of (a) a basic task and (b) a difficult task. Actual performance of an individual learning a basic task may display additional fluctuations between attempts (c).

Analysis of learning curves within the surgical education literature has been commonly been reported using traditional statistics. These statistical measures are best suited to comparing cohorts by splitting the groups and comparing performance with t-test, chi-square test or ANOVA (Cook et al., 2004; Joseph, Phillips, & Rupp, 2012). Alternatively, correlation statistics (eg. Pearson or Spearman) have also been used along with regression analysis. However, when delineating individual performance rather than grouped performance, cumulative sum analysis of learning curves (LC-CUSUM) and curve fitting are most commonly utilized (Biau, Williams, Schlup, Nizard, & Porcher, 2008; Marlies P. Schijven et al., 2004). The full text manuscript included in chapter 5 describes these analyses is detail in the discussion and outlines the reasons for selecting a novel analysis, namely k-means clustering, for this study.

2.6 Technical skill training curricula

2.6.1 Structuring curricula for technical skill training

Once the simulation models and assessment tools have been selected, a teaching model to deliver the curriculum is an important next step. Studies originating in the psychology literature and then adopted into surgical education have repeatedly demonstrated that a distributed practice training model, compared to a mass training design, is superior for learning (Mackay, Morgan, Datta, Chang, & Darzi, 2002; Moulton et al., 2006; Palter & Grantcharov, 2014).

The distributed practice model, also known as spaced repetition, breaks up the skills training into shorter time segments over multiples days or weeks. Massed practice is defined as continuous practice that occurs without rest between trials. The distributed practice learning paradigm was first reported in 1913 by a German psychologist Hermann Ebbinghaus, who studied this phenomenon in reference to memory recall and demonstrated that memorizing
random characters was more successful when separate in time (Ebbinghaus, 1913 (Reprinted Bristol: Thoemmes Press, 1999)). Subsequently, distributive practice has proven beneficial in procedural memory, learning fine and gross motor skills (Kwon, Kwon, & Lee, 2015; T. D. Lee & Genovese, 1989). Technical skills require both gross and fine motor movements. In the early 2000s, therefore, researchers compared distributed practice to mass practice in the area of surgical skills to assess whether there would be similar benefits.

The studies comparing mass practice to distributive practice were studied using simulated technical skills. Mackay et al compared massed practice to distributed practice on minimally invasive tasks and found a significant difference between the groups, favoring distributed practice (Mackay et al., 2002). Moulton et al compared the two practice methods for learning a microsurgery simulated bench task and found similar findings. Students completing distributed practice training outperformed massed practice when assessed with GRS and checklist scores (Moulton et al., 2006). Furthermore, the deliberate practice has also demonstrated successful transfer of skills learned in the simulation laboratory, to the real operating room (Palter & Grantcharov, 2014). Therefore, when I designing the technical skills training curriculum for detecting technical aptitude in medical students, a distributed practice-training model was used.

2.6.2 Mental practice as an adjunct for technical skill training

During surgical training, technical skills curricula using the models described above have gained tremendous momentum (Willis & Van Sickle, 2015). However, technical performance adjuncts that require minimal equipment and include focus on cognitive components are relatively new to the field of surgical education. Therefore mental practice (MP), as a performance adjunct for an advanced laparoscopic procedure, was studied in chapter 6 and was the first study in the surgical education literature to assess the effect of MP on advanced laparoscopic skills.
2.6.2.1 Definition and scope

MP is defined as the “cognitive rehearsal of a task without physical movement” or a “symbolic rehearsal of a physical action in the absence of any gross muscular movement (Richardson, 1969). Interchangeable terminology used to describe this cognitive process is: mental imagery (MI), mental rehearsal (MR) or visualization.

MP is performance adjunct that has been demonstrated to improve performance in many fields including sports, music and medical rehabilitation (Bar-Eli & Blumenstein, 2004; Nagano & Nagano, 2015; Spahn, 2015).

Mental practice as a performance enhancement technique was first reported in the early 1930s by Sackett (1934) and Perry (1939) who demonstrated that after MP, individual performance improved significantly for different psychology tasks (Perry, 1992; Sackett, 1934). Subsequently in the 1950s, MP began being used in sports when Soviet Union coaches used this technique to prepare their athletes for competitions (Ryba T, Stambulova N, & C., 2005). From the 1960s onwards, a growing interest in MP spurred researchers to conduct numerous studies to understand the effectiveness of this technique on motor performance. Feltz and Landers (1983) were the first to conduct a meta-analysis, including 60 studies comparing 146 effect sizes and reported that mentally practicing a motor skill is better than no practice at all but not as good as physical practice (D. L. Feltz & Landers, 1983). These results were again supported in a revised meta-analysis in 1988 (D.L. Feltz & Landers, 1988). Subsequently, Driskell et al. (1994) conducted a meta-analysis, which included 100 studies, and came to the same conclusion as their predecessors. In addition to the positive effect of mental practice on motor performance these authors studied the type of task, the retention internal from the MP intervention and the effect of the experience and concluded that: 1) the effect of MP is stronger for more cognitive tasks when compared to predominately physical tasks, 2) approximately 2 weeks after MP the effects have reduced to almost half and 3) MP is more effective for elite athletes as compared to novice athletes.
2.6.2.2 Mental practice theories

There are five theories that attempt to explain the positive effect of mental practice on performance. Each theory is outlined below.

1. Psychoneuromuscular theory

The psychoneuromuscular theory suggests that vivid, imagined events produce neuromuscular responses in the muscle fibers used when actually performing the task. Thus, the images produced in the brain during MP transmit impulses to the muscles for the execution of the imagined skill. In support of this theory, Jacobson recorded ocular movements in participants after visualizing images of the ‘Eiffel Tower’ or ‘this morning newspaper’ and consistently found increased activity during visualization as compared to relaxation (Jacobson, 1930). He then instructed subjects to imagine moving their arm and simultaneously held an electromyography (EMG) probe over the bicep. In visualizing the arm movement alone, the EMG probe over the bicep demonstrated increased muscular tension (Jacobson, 1931).

2. Symbolic learning theory

Symbolic learning theory differs from Psychoneuromuscular theory and states that MP is effective because individual actions are planned in advance and thus the individual benefits from being cognitively ready. Before a physical response is executed, advance planning optimizes motors sequence, sets task goals, and cognitively considers alternative solutions resulting in improved overall performance (Martin, Mortiz, & Hall, 1999).

3. Attention-arousal set theory

Attention and arousal set theory combines symbolic learning theory with the physiological aspects of psychoneuromuscular theory. According to this theory, imagery serves to improve performance physiologically, by helping the athlete to adjust his/her arousal level for optimal performance and cognitively by selectively attending to the task at hand and preventing distraction by irrelevant stimuli (R. S. Vealey, 1987).

4. Bio-informational theory
In 1979 Lang proposed that imagery is coded in a single, uniform, abstract manner in which the connections between these units play a specific and predetermine role. Lang’s model is unique because it emphasizes the importance of both input and output cues. He describes three main inputs: 1) image cue (e.g. written script), 2) image aid (e.g. film, script), 3) active participation in the imagined event. He then describes two output variables, 1) verbal report and 2) physiologic response which then feed back into the model to modify the image.

Figure 7 (Lang, 1979).

The reason for breaking up the imagery process into input and output cues is for research to focus on these components and to begin to study their impact on the overall process of MP and its effectiveness. For example, it is hypothesized that individuals who do not respond to MP therapy for obsessive compulsive behaviors only process the verbal cues rather than the complete interactive process (Lang, 1979).

Figure 7: Adapted from Lang’s model of input and output variables for emotional imagery.
5. **Ahsen’s triple code model of imagery (ISM)**

The triple code model has some similarities to Lang’s bioinformational theory in that it acknowledges the interaction of many processes that support and contribute to the success of MP. The triple code model simplifies the concept to focus on only three effects of the imagery. The effects are identified as ISM: I – the image itself, S-the somatic response, M – the meaning of the image to the individual. The incorporation of all three elements into imagery enhances overall performance and allows for performance coaches to focus on each component separately to optimize the mental imagery for the individual (Ahsen, 1984).

Combinations of these theories are generally accepted depending on the performance enhancement strategy being taught. In surgical education, the study of MP in relation to enhancing performance in sports is most commonly referenced and therefore understanding how athletes use MP as an adjunct to improve performance also clarifies how elements of MP have been incorporated into surgical training.

2.6.2.3 How athletes incorporate mental practice into training

It has become clear from historic and ongoing research that MP continues to demonstrate a positive effect on motor performance, including competitive sports. Studies from many competitive sports have demonstrated the positive effects of MP including diving, (Grouios, 1992) swimming, (Bar-Eli & Blumenstein, 2004) and golf (Brouziyne & Molinaro, 2005).

Three components are recommended for athletes to successfully incorporate MP into sport training: 1. beginning with a relaxation exercise, 2. vivid imagery using all five senses, 3. incorporating emotional imagery.

First, before engaging in MP, beginning with a relaxation exercise has been shown to improve performance (Wadey & Hanton, 2008). It is hypothesized that associating a state of control and calmness into mentally practicing a sport translates these associated feelings into true performance. Two common relaxation techniques are progressive muscle relaxations and deep breathing control. Progressive muscle relaxation involves maximally tensing the muscles throughout the body and systematically relaxing each muscle group until the body is completely
tranquil. Deep breathing control, also known as diaphragmatic breathing, involves breathing in deeply through the nose to expand the abdomen and then releasing each breath slowly through the mouth (Williams & Harris, 2006). Engaging in either technique prior to MP is recommended.

Second, the key to successfully using MP is being able to use visualization to produce an image that makes the individual feel like he/she is actually performing the sport (Hale, 1998; Holmes & Collins, 2001). Obtaining the highest level of vivid imagery requires performers to incorporate all five senses: sound, sight, touch, smell, and taste. Auditory imagery may be hearing the edge of an athlete’s ski as it cuts against the underlying snow. Visual imagery may be seeing a basketball move through the net of the hoop. Olfactory imagery may be the smell of chlorine before a swimmer enters the pool. Tactile imagery may be adjusting the grip before swinging a golf club. Kinesthetic imagery may involve feeling the optimal positions of movement before successfully executing a task (Holmes & Collins, 2001).

Third, in addition to the five senses, incorporating emotion into imagery has also been shown to strengthen the experience. Athletes may link emotions such as fear, anger, stress or anxiety to a particular performance. Recognizing and transforming negative emotions into positive feedback has proven beneficial (R. Vealey & Greenleaf, 2006). For example a soccer player may prepare for a penalty kick by thinking through the anxiety and pressure that is inevitably present during the execution of this task. They then are encouraged to visualize a successful goal and link it to the positive emotions associated with victory, such as elation, or pride (R. Vealey & Greenleaf, 2006).

However, all three of these components can be incorporated into MP that focuses on different aspects of enhanced performance. Recognizing the diversity of MP within sports, Hall et al. developed a classification system, separating imagery into five domains (Hall, Mack, Paivio, & Hausenblas, 1998; Short, Tenute, & Feltz, 2005).

1. Motivation-Specific (MS) – imaging focused on goal specific behavior such as winning a race or receiving a medal.

2. Motivational General-Mastery (MG-M) – imagery focused on coping during difficult performance moments, being mentally tough or working on confidence
3. Motivation General-Arousal (MG-A) – imagery that focuses on emotion in conjunction with sports competitions such as relaxation, anxiety, arousal.

4. Cognitive Specific (CS) – using imagery to practice and perfect sporting skills such as a double axel or slap shot.

5. Cognitive General (CG) – imagery of the strategies for competitive events such as a world cup soccer game.

Depending on the focus of the performance training, a different combination of the five domains above may be incorporated. Classically for top tear athletes preparing for a competition MS, MG-M and MG-A are the focus, whereas when novice players are learning a specific skill or task the focus of the imagery may be concentrated on CS and CG.

2.6.2.4 Mental practice in surgery

Similar to sports, surgery is a high-stakes environment with narrow performance margins and the potential for profound long-term consequences for patients being operated on. However, the application of MP in surgery is relatively new and has been focused primary on technical skill acquisition, focusing on domains CS and CG. In 2004 Sanders et al. were the first to study the impact of MP in basic open surgical skills. The authors had sixty-five medical students randomized to three groups receiving either physical practice alone or a combination of physical practice and mental rehearsal for placing simple interrupted sutures into a porcine foot model. The authors concluded that combining MP with physical practice resulted in equal performance scores as compared to continued physical practice alone (Sanders, Sadoski, Bramson, Wiprud, & Van Walsum, 2004). The same research group completed an additional study where students were randomly assigned to either a textbook description of an incision with simple interrupted suture closure, or MI of the same procedure. The MI group outperformed the textbook group (Sanders et al., 2008). Therefore, it became apparent that of the two cost effective modalities, MI demonstrated a positive impact on performance and could therefore be considered as an alternative to continued physical practice on more costly simulation models.
The impact of MI on minimally invasive techniques was studied first for endoscopy and for laparoscopy for junior level procedures. In 2009 Komesu et al. utilized the same methodology as Sanders et al. but compared gynecology residents’ performance on cystoscopy after being trained by either MP or a textbook description of the procedure. They concluded that the MP groups outperformed the textbook group by 15.1% (p=0.03). Initial study in the areas of laparoscopy demonstrated conflicting reports. Immenroth et al. demonstrated that MP did not improve laparoscopic cholecystectomy performance in already practicing surgeons (Immenroth et al., 2007) and Jungman et al. reported that MI did not show an advantage in performance of laparoscopic knot tying (Jungmann et al., 2011). However, Arora et al. criticized these studies and identified that an insufficient description of how MP was being delivered to the participants was problematic (S. Arora et al., 2010). Insufficient descriptive methodology made it difficult to understand the reasons for the discrepancies in the results and, furthermore, impossible to replicate these studies. To rectify this deficiency, Arora et al. developed and validated an MP protocol for surgery (Figure 8).
Figure 8: Mental Practice protocol for surgery, adopted from (S. Arora et al., 2010).

The MP protocol was tested during a randomized controlled trial (RCT) in novice surgical trainees performing laparoscopic cholecystectomy on a virtual reality simulator (S. Arora et al., 2010). The authors reported that the students randomized to the MP group outperformed the control group and scored statically superior despite baseline characteristics and technical skill levels being equal (S. Arora, Aggarwal, Sirimanna, et al., 2011). Therefore when conducting the RCT described in chapter 6 using MP as an adjunct for advanced laparoscopy in senior residents, MP was delivered to the participants using a similar protocol.
2.6.2.5 Mental practice script development

Using the protocol outlined in

Figure 8 requires the development of MP scripts, which in the setting of surgical education and acquisition of technical skills has focused primarily on CS and CG imagery, following three steps: 1. interview expert surgeons, 2. record the surgeon’s visual and tactile cues at each step of the operation, using both internal and external imagery, and 3. extraction of common themes from the expert script.

Step 1: Interview expert surgeons

It is generally accepted that the scripts should come from interviews of expert surgeons in the field, who perform the desired operation routinely. However, the optimal number of experts to be interviewed has not been defined. Arora et al. interviewed three expert surgeons but no rationale was reported to explain their sample size (S. Arora et al., 2010). Recognizing that collecting data through interviews is commonly done in qualitative research, it may be methodologically reasonable to stop collecting data when saturation has been met. Saturation is defined as the moment to stop interviewing new study participants ‘when the collection of new data does not shed any further light on the issue under investigation’ (Glaser.B. & Strauss.A., 1967). The sample size of experts to reach adequate saturation will likely differ depending on the complexity and predictability of the operative approach to the procedure selected. In the study outline in chapter 6, saturation was met after interviewing eight expert surgeons.

Step 2: Record the surgeon’s visual and tactile cues, using both internal and external imagery

When creating an MP script, it is recommended to incorporate four components: visual cues, kinesthetic cues, external and internal imagery, and each are explained in detail below.

As described in the sporting literature on MP, the scripts should incorporate as many of the five senses as possible in order to make the imagery realistic. In surgery, however, the imagery focused on visual and kinesthetic cues only because the laparoscopic surgical environment, from
a technical perspective, is less reliant on smell, taste and sound, thus these senses were excluded (Holmes & Collins, 2001).

Mental practice outcomes are improved when incorporating both external and internal imagery (Glisky, Williams, & Kihlstrom, 1996). External imagery is the ability to visualize a task from the third person perspective. For example, the expert being interviewed may reflect on common pitfalls performed by surgical residents when running the bowel, which would be verbalized in the third person, “they will grab the bowel and mesentery in the same bite, sometimes they will grab tiny bites of the bowel which increases the risk of injury if they slip off”. In contrast, internal imagery is when you see yourself execute the task in the first person. For example the expert being interviewed may verbalize, “I always check multiple times and from many angles that the stapler is perpendicular to the bowel. Once I am sure, I fire the stapler slowly”.

Step 3: Extraction of common themes from the expert script

The script is then compiled by extracting common themes from the transcribed document of each surgeon’s narration. The themes are generally organized into the steps of the operation and common visual and tactile cues extracted for each step, in both the first (internal imagery) and third (external imagery) person.

2.6.2.6 Mental practice as a strategy to decrease stress and improve non-technical skills

Apart from technical skills improvement, it has been shown that using MP in the simulated operative room can result in a significant reduction in the stress levels experienced by novice surgeons and thus improve non-technical performance (S. Arora, Aggarwal, Moran, et al., 2011).

The operating room is comprised of a team of professionals that work together to deliver safe surgical care to patients. The operating room is considered a high stakes, high stress environment, therefore the congruence of the team is essential to ensure performance is maintained despite the inherent stressors. The combined elements that comprise these human
factors within the operating room are termed non-technical skills (Yule & Paterson-Brown, 2012). Non-technical skills include: teamwork, communication, situational awareness and stress management and have been recognized as an important part of a functioning within the real operating room (Helmreich, 2000). Deficiencies in non-technical skill, especially communication within the operating room, have been identified as the reason for approximately 60% of perioperative complications (Greenberg et al., 2007). The physiologic release of cortisol induced by stress has been shown to modify cognitive processes in memory, performance and decision making (de Quervain, Roozendaal, Nitsch, McGaugh, & Hock, 2000; Johnston, Driskell, & Salas, 1997; Kirschbaum, Wolf, May, Wippich, & Hellhammer, 1996; Wolf, 2003). Therefore, utilizing MP as an adjunct to control stress may be a potential strategy to improve non-technical performance.

Four main valid and reliable assessment tools have been developed to evaluate non-technical skills of different team members within the operating room: surgeons, nurses or anesthetists (Table 7). These assessments are GRS that rely on direct observation of the operating room personnel and therefore have the same advantages and disadvantages to the technical skill GRS described in section 2.5.2.

Of the available tools, NOTSS and OSANTS both assess the surgeon role within the operative team. Given that the study described in chapter 6 assessed the effect of MP on the surgeon in the operating room, NOTSS was used because it was the only surgeon role assessment tool available at the time the study was completed.
<table>
<thead>
<tr>
<th>Name of assessment tool</th>
<th>Operating room team member being assessed</th>
<th>Non-technical skills included</th>
<th>Scale</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-technical skills for surgeons (NOTSS)</td>
<td>Surgeon</td>
<td>Situational awareness, Decision making, Communication and teamwork, Leadership</td>
<td>4 point Likert scale</td>
<td>(Yule et al., 2008)</td>
</tr>
<tr>
<td>Objective structured assessment of nontechnical skills (OSANTS)</td>
<td>Surgeon</td>
<td>Situational awareness, Decision making, Teamwork, Communication, Leading and directing, Professionalism, Managing and coordinating</td>
<td>5 point Likert scale</td>
<td>(Dedy et al., 2015)</td>
</tr>
<tr>
<td>Anaesthetists' non-technical skills (ANTS)</td>
<td>Anaesthetist</td>
<td>Situational awareness, Decision making, Task management, Team working</td>
<td>4 point Likert scale</td>
<td>(Lyk-Jensen, Jepsen, Spanager, Dieckmann, &amp; Ostergaard, 2014)</td>
</tr>
<tr>
<td>System for scrub practitioners' non-technical skills (SPLINTS system)</td>
<td>Nurses and technicians</td>
<td>Situational awareness, Communication and teamwork, Task management</td>
<td>4 point Likert scale</td>
<td>(Mitchell et al., 2013)</td>
</tr>
</tbody>
</table>
Chapter 3: Predictive value of background experiences and visual spatial ability testing on laparoscopic baseline performance among residents entering postgraduate surgical training

The systematic review (section 2.3.2) concluded that the visual spatial test PicSOR demonstrated a positive ability to predict laparoscopic skill on the VR simulators and laparoscopic box trainer. Therefore the following study was conducted and published as: Louridas M, Quinn E, Grantcharov TP. Predictive value of background experiences and visual spatial ability testing on laparoscopic baseline performance among residents entering postgraduate surgical training. Surg Endo.2015 June 20.

3.1 Abstract

**Background:** Emerging evidence suggests that despite dedicated practice, not all surgical trainees have the ability reach technical competency in minimally invasive techniques. While selecting residents that have the ability to reach technical competence is important, evidence to guide the incorporation of technical ability into selection processes is limited. Therefore, the purpose of the present study was to evaluate whether background experiences and 2D-3D visual spatial test results are predictive of baseline laparoscopic skill for the novice surgical trainee.

**Methods:** First year residents were studied. Demographic data and background surgical and non-surgical experiences were obtained using a questionnaire. Visual spatial ability was evaluated using the PicSOR, cube comparison (CC) and card rotation (CR) tests. Technical skill was assessed using the camera navigation (LCN) task and laparoscopic circle cut (LCC) task. Resident performance on these technical tasks was compared and correlated to the questionnaire and visual spatial findings.
Results: Previous experience in observing laparoscopic procedures was associated with significantly better LCN performance, and experience in navigating the laparoscopic camera was associated with significantly better LCC task results. Residents who scored higher on the cube comparison test demonstrated a more accurate LCN Path Length score \( r_{PL} = -0.36 \), \( p=0.03 \) and Angle Path \( r_{AP} = -0.426 \), \( p=0.01 \) score when completing the LCN task. No other significant correlations were found between the visual spatial tests (PicSOr, CC or CR) and LCC performance.

Conclusion: While identifying selection tests for incoming surgical trainees that predict technical skill performance is appealing, the surrogate markers evaluated correlate with specific metrics of surgical performance related to a single task but do not appear to reliably predict technical performance of different laparoscopic tasks. Predicting the acquisition of technical skills will require the development of a series of evidence-based tests that measure a number of innate abilities as well as their inherent interactions.

3.2 Introduction

Training programs that prepare candidates for performance in high-stakes, high pressure environments have two main opportunities to maximize the likelihood of graduates meeting their desired proficiency standards: 1) selecting appropriate candidates at the time of admission into training, and 2) optimizing the in-training learning environment. The importance of both strategies has been recognized in a number of high-stakes performance professions including aviation, sports and the armed forces (The Air Force, 2013; United States Coast Guard, 2013; Wickham & Dilworth, 1987). For example, the aviation industry relies on a battery of tests designed to objectively assess intelligence, coordination, instrument interpretation, general reasoning, and aptitude for understanding angles and bearings, before granting entry in the program (J. A. Bell, 1988). These tests are designed to select the students who are most likely to succeed on entering this difficult, high stress, resource heavy and costly training program (Holdsworth, 1988). Subsequently, a combination of didactic learning, simulation and real aircraft training methods are used to optimize learning for successful applicants.
Surgical practice is similarly a high-stakes, high-performance profession. While technical skill represents only a portion of the expertise expected of practicing surgeons, it is an unavoidable component of the field, and emerging evidence suggests that it may also be linked to patient outcomes (Birkmeyer et al., 2013). Furthermore, it is increasingly important given the changing training and practice landscape, including increasing reliance on more technically challenging minimally invasive techniques, and reduced work hour restrictions in the training milieu (Accreditation Council for Graduate Medical Education, 2013; Crothers, Gallagher, McClure, James, & McGuigan, 1999; Deziel et al., 1993). To account for these changes a number of innovations have been introduced to enhance the surgical training environment, including the use of validated simulation models and competency based training curricula ("CanMEDS 2015: The next evolution of the CanMEDS Framework," 2013; Palter, Orzech, Reznick, & Grantcharov, 2013; Royal College of Surgeons of England, 2013; Zevin B et al., 2013). Focusing on optimizing the selection process for surgical trainees will further complement these changes. While a range of subjective and objective selection criteria are used for admission into surgical training, these are generally limited to assessments of theoretical knowledge, clinical skills, and professionalism. However, it has been well documented that surgical trainees acquire technical skills at variable rates, and that some trainees may not reach competence despite dedicating time, effort and repetitive practice. Grantcharov, Schijven, and Alvand et al. have all reported this phenomenon, and have suggested that between 8 and 16 percent of contemporary surgical trainees fall within this group (A. Alvand, S. Auplish, H. Gill, & J. Rees, 2011; Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). Thus, while current criteria may effectively select candidates with desirable theoretical knowledge and professional attitudes, finding reliable measures to select residents able to reach competence in technical skills will undoubtedly further improve the selection process.

Several authors have suggested that previous experience with tasks requiring hand-eye coordination, and/or superior innate visual spatial 2D-3D conversion ability, may predict the acquisition of laparoscopic and endoscopic skills (A. G. Gallagher et al., 2003; D. Stefanidis, Korndorffer, et al., 2006). However, the evidence correlating these surrogate markers to surgical technical skills are limited and inconsistent (Maan, Maan, Darzi, & Aggarwal, 2012). Therefore, the purpose of the present study was to evaluate whether previous surgical experiences, non-
surgical experiences and 2D-3D visual spatial tests correlate with baseline laparoscopic skills in the novice surgical trainee.

3.3 Materials and Methods

3.3.1 Participants and setting

All individuals entering their first year of surgical training at the University of Toronto were considered eligible to participate in the present study, and were introduced to the study at the start of the Department’s mandatory surgical boot camp. Thirty-seven of a total of 57 eligible trainees consented to participate in the study. Trainees who declined to participate most commonly reported that they felt that laparoscopic skills were not pertinent to their surgical specialty, and were therefore not interested in the additional laparoscopic exposure provided by the study. The surgical boot camp is a mandatory introductory course for all first-year surgical trainees at the authors’ institution that provides teaching in basic surgical skills (e.g. sterile techniques,prepping and draping) and introductions to specialty-specific surgical techniques (e.g. laparoscopy and microsurgery). The boot camp takes place during the first two weeks of residency, and attempts to homogenize the varied exposures of the cohort during medical school. On completion of the boot camp a bell ringer examination is used to assess the trainees. Research Ethics Board approval was obtained prior to data collection, and all participants provided informed consent to participate in the study.

3.3.2 Demographics and participant questionnaire

The residents completed a demographics sheet along with a questionnaire quantifying surgical and non-surgical experiences thought to contribute to baseline surgical skill including exposure to laparoscopic procedures, video games and musical instruments (Adams, Margaron, & Kaplan, 2012; Ju, Chang, Buckley, & Wang, 2012; Paschold et al., 2011).
3.3.3 Visuospatial testing

Participants then completed three previously validated visual spatial tests designed to evaluate innate 2D-3D visual spatial ability. The intent was to combine visual spatial tests that have demonstrated positive relationships with minimally invasive technical skills in previous studies, and were therefore believed to measure the innate abilities necessary for incoming trainees to excel in laparoscopy (Buckley et al., 2013; Buckley et al., 2014; A. G. Gallagher et al., 2003; D. Stefanidis, Korndorffer Jr, et al., 2006). The following tests were used: the Pictorial Surface Orientation (PicSOr) test (A. G. Gallagher et al., 2003), the cube comparison test (CC), (Ekstrom, French, Harman, & Dermen, 1976) and the card rotation test (CR) (Ekstrom et al., 1976). The PicSOr test evaluates 2D-3D conversion ability by requiring participants to orient a rotating arrow at a 90-degree angle to one side of an underlying cube. Initially, students used the PicSOr test in practice mode, receiving immediate feedback from the software concerning the actual arrow to cube angle. No limitations were set for the duration of practice, and a single instructor was available to answer questions during this time. Once participants felt ready to proceed, the software was switched to experiment mode (Figure 9). During the experiment, 35 angle estimations were completed and students were instructed to complete the task as quickly and accurately as possible. No time restriction was enforced. The scores were graded as described by the creators of PicSOr (A. G. Gallagher et al., 2003). The students then completed the CC tests and CR test paper tests, administered as directed by the test copyright holder’s (Educational Testing Service, Princeton New Jersey) official test administration manual Figure 10 (Ekstrom et al., 1976). The participants started by performing three practice exercises before starting the test, with a single instructor present to correct and explain the questions. Next, the two-page test was administered and timed, with 3 minutes allotted per page, totaling 42 cube comparison exercises and 20 rows of card rotation tests. The students were instructed to work through the questions as quickly and accurately as possible. The tests were graded with the answer key provided in the test administration manual, with negative marking for incorrect answers.
Figure 9: Pictorial Surface orientation test (PicSOR) used to assess 2D-3D perception ability. a. Setup to change between practice and experiment mode. b. rotating arrow oriented to lie 90 degrees to the underlying cube.

Figure 10: Paper tests used to assess 2D-3D visual spatial ability included the (a) card rotation test (CR) and (b) cube comparison test (CC)
3.3.4 Laparoscopic skills familiarity session

Residents were exposed to basic laparoscopic techniques during a mandatory 2 hour boot camp session. The purpose of this session was to ensure that all residents were familiar with the laparoscopic instruments and basic simulated tasks. First, students watched a video illustrating each of the three Fundamentals of Laparoscopic Surgery (FLS) exercises. Subsequently, students attempted the three tasks: the peg transfer task, the laparoscopic circle cut (LCC) task, and intracorporeal knot tying. (Peters et al., 2004) Additionally, residents had the opportunity to attempt three basic LapSim virtual reality simulator (Surgical Science, Gothenburg, Sweden) tasks including: laparoscopic camera navigation (LCN), the coordination task, and the lifting and grasping task. Three surgical faculty and two senior surgical residents were present at the laparoscopic session to answer questions and demonstrate the tasks for the first year trainees. No formal training was offered during this session.

3.3.5 Assessment of laparoscopic baseline skill

Students were given a maximum of 300 seconds to complete the laparoscopic circle cut task in the box trainer. Students were asked to cut out the circle as quickly and accurately as possible (E. M. Ritter & Scott, 2007). The circle cut task was scored by a single grader using the objective FLS scoring system, with lower scores representing better performance: final score (max 300)= total time (seconds) + error (surface area of white gauze from the black line). Each resident was then given three attempts at the LCN task on the virtual reality (VR) simulator and the best score was retained for assessment. The LCN task was scored using three VR metrics: total time, instrument path length and instrument angular path. LCN total score was generating by adding these three components in equal weights.

3.3.6 Statistical Methods

One-way analysis of variance (ANOVA) was used to assess the relationships between previous surgical and non-surgical experiences and the laparoscopic tasks. Hochberg post hoc analysis
was used in the cases where the ANOVA was significant in order to identify the level at which significance occurred. Leven’s test was met for all ANOVA calculations. Spearman’s rho correlation analysis was used to examine the relationship between the three innate visual spatial ability tests and the two laparoscopic skill test scores, because of non-normal distribution of visual spatial test scores. Values are presented as median (range) unless stated otherwise. All data was analyzes using SPSS 22.0 software (SPSS Inc., Chicago, Illinois).

3.4 Results

3.4.1 Demographics

The group included 37 first year surgical residents, 27 men and 11 females, all right handed, with a median age of 27 years of age (range 23-37 years). Seven surgical specialties were represented in the sample (14 from general surgery, 8 from orthopaedic surgery, 4 from urology, 3 from plastic surgery, 3 from vascular surgery, and one each from 5 different specialties). Thirty-three participants had no previous surgical training outside of medical school, 3 had completed 1-2 junior resident years of surgery training, and 1 had completed a full surgical residency abroad and fellowship training in vascular surgery in Canada, before being accepted into Canadian surgical residency training. None of the participants had completed the FLS standardized course (Table 8).
<table>
<thead>
<tr>
<th>Background characteristic</th>
<th>No. of participants</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>26</td>
</tr>
<tr>
<td>Females</td>
<td>11</td>
</tr>
<tr>
<td><strong>Handedness</strong></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>37</td>
</tr>
<tr>
<td>Left</td>
<td>0</td>
</tr>
<tr>
<td><strong>Surgical program</strong></td>
<td></td>
</tr>
<tr>
<td>Plastic surgery</td>
<td>3</td>
</tr>
<tr>
<td>General surgery</td>
<td>14</td>
</tr>
<tr>
<td>Ear nose and throat surgery</td>
<td>1</td>
</tr>
<tr>
<td>Orthopaedic surgery</td>
<td>8</td>
</tr>
<tr>
<td>Urology</td>
<td>4</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>2</td>
</tr>
<tr>
<td>Cardiac surgery</td>
<td>2</td>
</tr>
<tr>
<td>Vascular surgery</td>
<td>3</td>
</tr>
<tr>
<td><strong>Previous surgical training</strong></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>33</td>
</tr>
<tr>
<td>1-2 junior years</td>
<td>3</td>
</tr>
<tr>
<td>Completed vascular surgery</td>
<td>1</td>
</tr>
<tr>
<td><strong>Fundamental of laparoscopic surgery</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>37</td>
</tr>
</tbody>
</table>
Table 9: Previous surgical and non-surgical experiences

<table>
<thead>
<tr>
<th>Type of previous experience</th>
<th>No. of participants</th>
<th>LCN p - values</th>
<th>LCC p - values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed laparoscopic procedures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0</td>
<td>0.04*</td>
<td>0.55</td>
</tr>
<tr>
<td>&lt;10</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-20*</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt;20*</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drove the laparoscopic camera</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>2</td>
<td>0.55</td>
<td>0.01*</td>
</tr>
<tr>
<td>&lt;10</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-20</td>
<td>6</td>
<td>0.55</td>
<td>0.01*</td>
</tr>
<tr>
<td>&gt;20*</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Musical instrument</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>13</td>
<td>0.86</td>
<td>0.70</td>
</tr>
<tr>
<td>Piano</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>String instrument</td>
<td>4</td>
<td>0.86</td>
<td>0.70</td>
</tr>
<tr>
<td>Wind instrument</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Musical royal conservatory grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>19</td>
<td>0.37</td>
<td>0.14</td>
</tr>
<tr>
<td>&lt; grade 5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; grade 5</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exposure to video games</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>17</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>&gt;1-2 times/week</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 times/month</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-2 times/year</td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Indicates level of significance after post-hoc analysis with Hochberg for alpha <0.05
NB: LCN - laparoscopic camera navigation, LCC - laparoscopic circle cut
3.4.2 Participant questionnaire

Of all the baseline surgical and non-surgical experiences assessed, only previous laparoscopic experience was significantly predictive of baseline surgical skill on entering residency. Students who had observed >10 laparoscopic procedures scored significantly higher on the LCN task (mean 63.0 points compared to 57.9 points; p=0.04), when compared to those who had observed between 0 and 10 procedures. Furthermore students who reported laparoscopic camera navigation experience in more than 20 operative cases performed the LCC task significantly faster and with more accuracy (lower score) than students who had less laparoscopic experience (means score of 240.3 points compared to 323.8 points; p =0.01). No relationship was seen between performance on either technical task and video game or previous music instrument experience, whether assessed by instrument played or Royal Conservatory of Music grade level completed (Table 9).

3.4.3 Correlation of visual spatial skills and laparoscopic baseline performance

Residents who scored higher on the CC test demonstrated more accurate LCN path length ($r_{s(PL)} =-0.36$, p=0.03) and angle path ($r_{s(AP)} =-0.426$, p=0.01) scores during the LCN task. However, participants’ time to complete the LCN task was not significantly associated with their CC test scores ($r_{s(time)} =-0.04$, p=0.84). No significant correlations were identified between LCN metrics and PicSOr ($r_{s(PL)} =0.19$, p=0.25, $r_{s(AP)} =0.25$, p=0.141, $r_{s(time)} =-0.27$, p=0.14) or CR ($r_{s(PL)} =-0.06$, p=0.74, $r_{s(AP)} =-0.06$, p=0.74, $r_{s(time)} =-0.07$, p=0.70) test scores. In addition, no significant correlations were observed between LCC time, error and total score when compared to all the three visual spatial tests (Table 10).

When examining the relationship between individual visual spatial tests a significant correlation was seen between CC and CR tests ($r_s =0.33$, p=0.05), suggesting that these tests measure related visual spatial abilities. However, no correlation was seen between the PicSOr and CC or CR tests ($r_s =-0.01$, p=0.93 and $r_s =-0.07$, p=0.69 respectively).
Table 10: Correlation of 2D-3D innate ability tests with laparoscopic surgical skill

<table>
<thead>
<tr>
<th>Name of technical task</th>
<th>PicSOr</th>
<th>p - value</th>
<th>Card rotation test</th>
<th>p - value</th>
<th>Cube comparison test</th>
<th>p - value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Laparoscopic camera</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>navigation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>0.27</td>
<td>0.11</td>
<td>-0.07</td>
<td>0.70</td>
<td>0.04</td>
<td>0.84</td>
</tr>
<tr>
<td>Path Length</td>
<td>0.19</td>
<td>0.25</td>
<td>-0.06</td>
<td>0.74</td>
<td>-0.36</td>
<td>0.03*</td>
</tr>
<tr>
<td>Angular Path</td>
<td>0.25</td>
<td>0.14</td>
<td>-0.07</td>
<td>0.66</td>
<td>-0.43</td>
<td>0.009*</td>
</tr>
<tr>
<td>Combined Score</td>
<td>0.22</td>
<td>0.19</td>
<td>-0.08</td>
<td>0.64</td>
<td>-0.42</td>
<td>0.01*</td>
</tr>
<tr>
<td><strong>Laparoscopic circle cut</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>-0.08</td>
<td>0.66</td>
<td>0.009</td>
<td>0.96</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>Error</td>
<td>0.13</td>
<td>0.44</td>
<td>-0.18</td>
<td>0.29</td>
<td>0.11</td>
<td>0.52</td>
</tr>
<tr>
<td>Combined score</td>
<td>-0.05</td>
<td>0.77</td>
<td>-0.16</td>
<td>0.35</td>
<td>0.14</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Indicates statistical significance for an alpha <0.05

3.5 Discussion

Recent evidence has challenged the traditional assumption that all trainees will become technically component surgeons given dedicated time and practice. Incorporating a measure of technical aptitude into the existing selection process to identify candidates who are more likely to reach technical competency during training is not only appealing, but could be relatively easy to implement. However, despite investigating a broad range of potentially predictive background characteristics and visual spatial tests, only previous laparoscopic experience seemed to correlate with baseline simulated laparoscopic performance.

Few studies have investigated the relationships between previous surgical experience and baseline laparoscopic skills in novice trainees. In a study of 25 medical students, Benarjee et al. found that those who had over 20 hours of experience assisting with laparoscopic operations had
significantly faster performance on an intracorporeal knot tying task (Banerjee, Cosentino, Hatzmann, & Noe, 2010a; Buckley et al., 2014). Similarly, in the present study, residents who had previously navigated a laparoscopic camera in more than 20 cases scored better on the LCC task compared to trainees with less experience. Additionally, those who had observed more than 10 hours of laparoscopic surgery performed significantly better on the LCN task. These findings suggest that exposure to laparoscopy during medical school may be one way to gauge initial performance on basic laparoscopic tasks.

Other non-surgical characteristics that can be obtained through a questionnaire such as: gender, handedness, typing performance, musical instrument ability, sewing, self-perceived dexterity, and interest in surgery and sports, have repeatedly been shown to not be predictive of surgical skill in a large number of studies including the present study (Cope & Fenton-Lee, 2008; Lokuge, 2012; Madan et al., 2008). However, in contrast to the present study, increased video game experience has been previously shown to correlate with superior baseline laparoscopic performance (Hislop et al., 2006; Nomura et al., 2008; Paschold et al., 2011; Van Hove et al., 2008). Paschold et al. reported that video game experience correlated with better VR performance in two tasks (grasping and retracting tissue, and applying clips on the cystic duct and artery) (Paschold et al., 2011). In the present study, the low number of participants in each video game subgroup may explain the discrepancy in these findings. Nevertheless, even when gamers exhibit an initial advantage in performance, Van Hove et al. reported that their performance equalizes with that of non-gamers following even modest amounts of practice (Van Hove et al., 2008). If the technical gaming advantage can be quickly learned, quantifying gaming experience is of little utility for surgical selection. Therefore, while non-surgical information obtained through questionnaires or curricula vitae, may have some value in highlighting extracurricular accomplishments and interests among candidates for surgical training, this information does not appear to predict initial technical performance in the contemporary training milieu.

While it is generally accepted that visual spatial 2D-3D conversion ability is required to perform laparoscopic and endoscopic procedures, there is considerable variability in the reported correlations between the results of formal visual spatial testing and performance on these procedures. As a result, the reliability of formal visual spatial testing is questionable. In the present study, residents who obtained a higher score on the CC test performed better on the VR
LCN task in two of three metrics, but not on the LCC box-training task. Furthermore, despite a significant relationship between participant’s performance on the CR and CC task, no correlation was identified between CR test scores and performance on either of the studied laparoscopic tasks. In keeping with these inconsistencies, Enochsson et al. found that the CR test did not reliably predict simulated gastroscopy performance in the novice trainee. Additionally, the authors found that expert gastroscopists performed worse on the CR test when compared to novice trainees (L. Enochsson et al., 2006). In contrast to these findings, Arora et al. reported a positive correlation between medical students’ CC, CR and PicSOr test results, and their performance on simulated endoscopic sinus surgery (H. Arora et al., 2005; Westman et al., 2006). The PicSOr test, while designed specifically to test the perceptual ability thought to be required for laparoscopy, has also demonstrated inconsistent associations with surgical performance. Gallagher et al. reported positive correlations between surgical trainees’ PicSOr results and their performance on the LCC task, and Kolozsvari reported similarly positive correlations with laparoscopic peg transfer performance in the novice trainee. In contrast Stefanidis et al. demonstrated no correlation between PicSOr test results and a range of laparoscopic simulator tasks (D. Stefanidis, Korndorffer, et al., 2006). This is in keeping with the results of the present study, where the PicSOr test failed to predict performance on any simulated laparoscopic tasks. Similarly Thus, it would appear that although visual spatial ability is thought it be essential for laparoscopic performance, the performance scores generated by the PicSOR, CC and CR tests do not appear to accurately and consistently predict baseline technical skill performance in novice trainees entering surgical residency. As a result, these tests in their current form do not appear to be sufficiently reliable for use during trainee selection.

We acknowledge several limitations to the present study. First, due to the structure of the final bell ringer examination, there was only enough time for a single attempt at the LCC. Recognizing that baseline skill assessment can vary substantially in novice trainees between consecutive attempts, using either the best or average score of three trials may have improved the accuracy of assessment by reducing the effect of test-retest variance. Second, we were limited by the relatively small size of our study cohort, potentially increasing the risk of a type II error in our statistical analyses, most notably with respect to comparisons made using ANOVA. Finally, it is possible that participants who were entering surgical specialties with limited or no laparoscopic procedures may have been less motivated to perform well, biasing the results.
However, it would appear that many of these residents declined to participate in the study in the first place, given the large number of non-participants who cited this reason. Nevertheless, despite these limitations, we believe that the concurrent assessment of multiple different potential predictors of baseline laparoscopic skills in novice trainees in our study provides valuable, novel evidence concerning the reliability of these assessment tools.

The findings of the present study suggest that trainees who have previously had the opportunity to observe and participate in laparoscopic cases have better baseline laparoscopic skills. However, surgical technical skills, and the aptitude to acquire them, do not appear to transfer from other life exposures or non-surgical experiences. Furthermore, formal tests of visual spatial ability have demonstrated inconsistent associations with technical skills. The execution of any given technical task is likely to be the result of a complex matrix of aptitudes that combine in different ways and result in varied performance levels. Therefore, isolated formal tests of visual spatial ability likely oversimplify this matrix of aptitudes, resulting in the previously described inconsistent relationships with technical performance. Given these findings, instead of relying on surrogate markers, future work may be best directed toward assessing the predictive value of performance on higher-fidelity tasks that more closely replicate actual surgical tasks. Examples might include performance of tasks in the simulation laboratory that have been validated and shown to transfer into the operating room.

While the use of selection tests for incoming surgical trainees that predict future technical skill performance would be beneficial in optimizing the technical competence of graduating surgeons, surrogate markers such as non-surgical experience and visual spatial tests do not appear to be reliable predictors. Given the poor predictive ability of the range of factors and tests evaluated in the present study, and the conflicting data in the literature, it is reasonable to conclude that novel approaches are required to identify reliable predictors of technical skill performance in surgical trainees, and to identify those candidates most likely to reach technical competence in the contemporary training environment.
Chapter 4: Optimizing the selection of general surgery residents - a national consensus

The content of this chapter has been submitted for review and has been peer reviewed by the American College of Surgeons Accredited Education Institute.

4.1 Abstract

**Background:** Surgical programs strive to recruit trainees who will graduate as competent surgeons, however selection processes vary between institutions. The purpose of the present study was to 1) identify the current Canadian general surgery (GS) selection processes 2) solicit program directors’ (PDs) opinions on the proportion of trainees who have difficulty achieving competence and 3) establish consensus on the desired attributes of GS candidates, and the technical skills that would be most indicative of future performance.

**Methods:** Delphi consensus methodology was used. An open-ended, followed by a closed-ended questionnaire, formulated as a 5 point Likert scale, was administered. A Cronbach’s alpha of $\geq 0.8$ with 80% of responses in agreement (4 - agree and 5 - strongly agree) determined the threshold for consensus.

**Results:** The first and second rounds were completed by 14 and 11, of a potential 17, GS PDs, respectively. Only 2 programs reported assessing technical skill, however, the majority of PDs felt that 5-15% of residents had difficulty reaching competence in this area. Consensus was excellent (alpha = 0.92). The top traits for success in GS included: work ethic, passion for surgery. Technical skills felt to be most appropriate were open tasks (one-handed tie and subcuticular suture) and laparoscopic tasks (coordination, grasping and cutting).

**Conclusion:** PDs indicate that 5-15% of graduating residents had difficulty reaching competence in technical skill. Despite this, objectively assessing this domain is rarely included in the selection process. Consensus among PDs suggests that basic open and laparoscopic skills are appropriate for inclusion into the selection process.
4.1.1 Background

Surgical programs strive to recruit trainees who will graduate as competent surgeons. To structure entry into surgical post-graduate training programs in North America, national match systems are used to pair final year medical students to specialty programs (e.g. general surgery (GS) or neurosurgery) (CaRMS, 2016; NRMP, 2016). Therefore, unlike many countries, medical students are admitted directly into a surgical specialty program without completing an internship or advancing from basic to advanced surgical training (Condon et al., 2013; Erasmus, 2012; Anthony G. Gallagher, Leonard, & Traynor, 2009; Anthony G. Gallagher et al., 2008; Levitt & Klein, 1991). Thus, Canadian and American programs are in a unique position because they are selecting candidates into specialty, without having the opportunity to assess their independent performance in the clinical environment or their acquisition of technical skill in the operating room.

Intuitively, however, students who apply to enter surgical training likely enjoy working with their hands and may self-select as better technicians. It has been reported that students who apply to surgical specialties have a higher self-perceived confidence in their manual dexterity and ability to “work well with their hands,” as compared to their medical colleagues (J. Y. Lee, Kerbl, McDougall, & Mucksavage, 2012; Van Hove et al., 2008). However, when comparing these two groups with objective technical skill assessment metrics, the incoming surgical trainees do not outperform the internist (Cope & Fenton-Lee, 2008; Panait et al., 2011). Self-selection cannot be relied upon to ensure that surgical applicants have a high potential for technical performance and therefore it may be appropriate that surgical programs are given the responsibility to make this assessment instead.

In the current North American system, technical skill is not routinely a component of the selecting process. This may be due to the strong belief, supported by Ericsson learning theory, that ongoing practice and mentorship will eventually translate into expert performance (K. A. Ericsson, 2007). However, with work-hour restrictions, increasing complexity of surgical techniques and increased patient safety concerns, the feasibility of this model has been challenged (Kothari & Ponce, 2014). It has been reported that trainees are not reaching their expected technical milestones by the end of training, which are then reflecting in their performance at the fellowship level (Antiel, Thompson, Camp, Thompson, & Farley, 2012).
United States fellowship PDs reported that a significant proportion of GS fellows could not independently perform a laparoscopic cholecystectomy or operate unsupervised for more than 30 minutes during a major procedure (Mattar et al., 2013). Given these reports, it may be beneficial for training programs to adjust the GS selection process to recruit applicants who are able reach technical competence within the restrictions of the current training environment. However, there is a lack of evidence to guide this aspect of the selection process (Kenny, McInnes, & Singh, 2013).

Therefore, the purpose of the present study was to: 1) identify the current components used in the GS selection process at different institutions; 2) solicit program directors’ opinions on the proportion of trainees who do not achieve the minimum standards expected of graduating trainees; and 3) establish national consensus on the desired attributes of GS candidates, and the technical skills that would be most indicative of future performance.

4.2 Methods

Research ethics

The University of Toronto Ethics Review Board approved this study.

4.2.1 Current selection practices

All Canadian GS program directors (PDs) were invited to participate. In Canada, all training programs are structured under the umbrella of a University with a recognized medical school, in contrast to individual hospital programs. An online questionnaire, administered using Survey Monkey (Palo Alto, CA), was used to identify the current components used in the GS selection process across the country. Although the written application is standardized by the national match system (CaRMS- Canadian Resident Matching Service), PDs were asked to provide the weighted score for each component of the application at their institution. In addition, PDs were asked whether applicants’ clinical knowledge, decision-making and technical skill were evaluated during the selection process, and what percent of trainees they felt had difficulty reaching competence in these three domains by the time of graduation.
4.2.2 Delphi consensus methodology

A Delphi questionnaire was administered to gain consensus on which candidate-specific attributes are important for residents to succeed in GS training. In addition, consensus was sought on the simulated technical skills (both open and laparoscopic) that are most likely to be indicative of a trainee’s aptitude to acquire more complex surgical skills and thus future performance.

The Delphi methodology was originally developed in the 1950s by the RAND Corporation to evaluate trends in technology on warfare, but continues to be widely used to create public policy, clinical guidelines or to formulate training recommendations by aggregating the opinions of experts, where little empirical evidence is available (Elissen, Struijs, Baan, & Ruwaard, 2015; Loeffen et al., 2015; RAND, 1976; B. Zevin, Levy Js Fau - Satava, Satava Rm Fau - Grantcharov, & Grantcharov). This methodology is comprised of four essential components: an expert panel, the promotion of anonymous responses, multiple rounds of questions, and statistical feedback to encourage convergence of responses until an acceptable consensus is met (RAND, 1976).

4.2.3 Expert panel

Canadian general surgery PDs were invited by email to participate in this Delphi process. This group of individuals was selected to participate due to their unique expertise with trainee selection, as acquired through leadership in this process at their respective institutions.

4.2.4 Anonymity

A strength of the Delphi technique is that it protects against bias by prohibiting face-to-face contact amongst the panel members, thus decreasing dominant verbal opinions, seniority or in-person arguments which have been reported to sway the panel (Cuhls, 2005; Murphy et al.,
In the present study, individual anonymous opinions were encouraged through an online questionnaire, limiting the risk of interaction between panel members (Murphy et al., 1998).

### 4.2.5 Rounds of questions

The Delphi process calls for a minimum of two rounds of questions, with the first open-ended and the second closed-ended (Cuhls, 2005; Powell, 2002). Open-ended questions encourage responses from the expert panel without directing their opinions to multiple-choice answers (Powell, 2002). In the present study, the first round of open-ended questions was then supplemented with literature in the field to ensure completeness for the second-round of closed-ended questions. In the second round, responses to closed-ended questions are solicited using a Likert scale to allow for statistical feedback to panelists, and to encourage convergence of responses to create consensus (Jairath & Weinstein, 1994). In our study, the closed-ended questions were formulated on a 5-point Likert scale (1 - strongly disagree, 2 – disagree, 3 - somewhat agree, 4 – agree, 5 - strongly agree).

### 4.2.6 Consensus

Consensus for each section of the Delphi is calculated with a Cronbach’s alpha, which is a statistical measure of internal consistency or homogeneity of expert responses. A Cronbach’s alpha of $\geq 0.8$ has previously been reported to be an acceptable benchmark for consensus and was therefore used as the cutoff for this study (Graham, Regehr, & Wright, 2003). Subsequent rounds are performed until consensus is met. The present study met consensus after the second round.

### 4.2.7 Recommendations for incorporation into selection

Once consensus was achieved, each questionnaire item was assessed for whether the agreement among experts was found to be positive, neutral or negative. Only items that reached positive
agreement were recommended for incorporation into the selection process. Positive agreement was achieved when >80% of experts rated the attribute or skill as a 4 (agree) or 5 (strongly agree) on the Likert scale. Negative agreement was established when >80% of responses were either 1 (strongly disagree) or 2 (disagree) on the Likert scale, opposing inclusion of those attributes or skills into selection. Neutral agreement included all other responses (B Zevin, Levy, Satava, & Grantcharov, 2012).

4.3 Results

4.3.1 Current selection process

Of the 17 GS program directors across Canada, 14 and 11 participated in the first and second rounds of the questionnaire, respectively. Excellent representation was achieved across the country, with at least one GS program responding from each province (Table 11). The selection process was consistently reported to occur in two stages: review of written applications, followed by an in-person interview. All programs reported using a structured review process for the written application, however the weighted score of the components varied between programs: personal statement 5-25%, curriculum vitae 10-35%, research involvement 10-15%, perceived dedication to GS 5-20%, and reference letters 10-40%. Candidate’s application scores were then ranked and in-person interviews offered to those candidates who achieved scores above the program specific cutoff. Following the interview, each applicant’s final rank score was determined by combining the written application score (weighted 25-60%) and the interview score (weighted 40-75%). Ten of 14 programs reported using reference letters and scripted interview scenarios to assist with decision making at the time of candidate selection. Six of 14 assess clinical knowledge using transcripts, reference letters, curriculum vitae and interview scenarios. Finally, only 2 of 14 programs assess technical skill through simulated stations and reference letters.

The majority of PDs felt that less than 5% of trainees have difficulty reaching competence in clinical knowledge (range 0-10%), that 5-10% of trainees has difficulty in decision-making
(range 0-20%), and that 5-15% (range 0-15%) in technical skill by the time of completion of training (Table 12).

Table 11: Participating Canadian General Surgery Programs

<table>
<thead>
<tr>
<th>Participating Programs (n=14)</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>McMaster University*</td>
<td>Hamilton, Ontario</td>
</tr>
<tr>
<td>University of Montreal*</td>
<td>Montreal, Quebec</td>
</tr>
<tr>
<td>University of Manitoba*</td>
<td>Winnipeg, Manitoba</td>
</tr>
<tr>
<td>Dalhousie University</td>
<td>Halifax, Nova Scotia</td>
</tr>
<tr>
<td>Memorial University</td>
<td>St. John, Newfoundland</td>
</tr>
<tr>
<td>Sherbrooke University</td>
<td>Montreal, Quebec</td>
</tr>
<tr>
<td>University of Alberta</td>
<td>Edmonton, Alberta</td>
</tr>
<tr>
<td>University of British Columbia</td>
<td>Vancouver, British Columbia</td>
</tr>
<tr>
<td>University of Calgary</td>
<td>Calgary, Alberta</td>
</tr>
<tr>
<td>University of Ottawa</td>
<td>Ottawa, Ontario</td>
</tr>
<tr>
<td>Queens University</td>
<td>Kingston, Ontario</td>
</tr>
<tr>
<td>University of Saskatchewan</td>
<td>Saskatoon, Saskatchewan</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>Toronto, Ontario</td>
</tr>
<tr>
<td>Western University</td>
<td>London, Ontario</td>
</tr>
</tbody>
</table>
* Programs that only participated in the first round

4.3.2 Delphi consensus methodology

The overall consensus for the questionnaire was excellent, with a Cronbach’s alpha of 0.92. Internal consensus was reached for each of the four sections of the questionnaire, namely: 1) desired candidate attributes, 2) open surgical skills, 3) virtual reality (VR) simulation skills and 4) laparoscopic box training skills, with a Cronbach’s alpha of 0.87, 0.96, 0.92 and 0.80 respectively.
Table 12: General Surgery program director's responses to clinical knowledge, decision-making and technical skill during selection and at the time of graduation

<table>
<thead>
<tr>
<th>Area</th>
<th>During selection</th>
<th>At the time of graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Clinical knowledge</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Decision making</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Technical skill</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>
Table 13: Desired candidate attributes for selection into General Surgery

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Cronbach's alpha = 0.87</th>
<th>Median (IQR)</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work ethic</td>
<td></td>
<td>5 (5,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Passion for General Surgery</td>
<td></td>
<td>5 (5,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Professionalism</td>
<td></td>
<td>5 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Ability to work in a team</td>
<td></td>
<td>5 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Sound judgment</td>
<td></td>
<td>5 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Ability to make decisions</td>
<td></td>
<td>5 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Assimilates information to formulate an opinion</td>
<td></td>
<td>5 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Independence</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Ability to multitask</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Technical skill</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Ability to accept criticism</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Ability to think quick on his/her feet</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Collaborative skills</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Cool under pressure</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Humility to learn</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Communication</td>
<td></td>
<td>4 (4,5)</td>
<td>positive</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td>4 (4,4)</td>
<td>positive</td>
</tr>
<tr>
<td>Trainability</td>
<td></td>
<td>4 (3,4)</td>
<td>neutral</td>
</tr>
<tr>
<td>Leadership</td>
<td></td>
<td>4 (3,4)</td>
<td>neutral</td>
</tr>
</tbody>
</table>

IQR: interquartile range
Table 14: Appropriate simulated surgical skills for selection into General Surgery

<table>
<thead>
<tr>
<th>Open Skills</th>
<th>Median (IQR)</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach's alpha = 0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*One handed ties</td>
<td>4 (4,5)</td>
<td>Positive</td>
</tr>
<tr>
<td>*Interrupted subcuticular suturing</td>
<td>4 (4,4)</td>
<td>Positive</td>
</tr>
<tr>
<td>*Running subcuticular suturing</td>
<td>4 (4,4)</td>
<td>Positive</td>
</tr>
<tr>
<td>Horizontal mattress</td>
<td>4 (4,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Vertical mattress</td>
<td>4 (4,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Nevus removal</td>
<td>4 (4,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Simple interrupted suture</td>
<td>4 (3,5)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Two handed tie</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Hand sewn bowel anastomosis</td>
<td>4 (2,5)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Chest tube insertion</td>
<td>3 (3,5)</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Virtual Reality Simulation Skills

<table>
<thead>
<tr>
<th>Virtual Reality Simulation Skills</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cronbach's alpha = 0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>*Coordination</td>
<td>4 (4,4)</td>
<td>Positive</td>
</tr>
<tr>
<td>*Grasping</td>
<td>4 (4,4)</td>
<td>Positive</td>
</tr>
<tr>
<td>*Cutting</td>
<td>4 (4,4)</td>
<td>Positive</td>
</tr>
<tr>
<td>Camera navigation</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Instrument navigation</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Skill</td>
<td>Score</td>
<td>Difficulty</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-------</td>
<td>------------</td>
</tr>
<tr>
<td>Clip applying</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Lifting and grasping</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Bowel handling</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Precision Task</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Hand eye coordination</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Fine dissection</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Catheter insertion</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Colonoscopy</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Seal and cut</td>
<td>3 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Gastroscopy</td>
<td>3 (2,4)</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

**Laparoscopic Box Trainer Skills**

<table>
<thead>
<tr>
<th>Skill</th>
<th>Score</th>
<th>Difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peg transfer</td>
<td>3 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Laparoscopic circle cut</td>
<td>3 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Extracorporeal Knot tying</td>
<td>3 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Endoloop placement</td>
<td>3 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Intracorporeal knot tying</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Camera navigation</td>
<td>4 (3,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Running intracorporeal suture</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
<tr>
<td>Cholecystectomy on porcine bowel model</td>
<td>4 (2,4)</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

* inclusion into selection of technical skill; IQR: interquartile range
Sixteen of the 18 individual personality attributes met the criteria for inclusion into the selection processes. The top 7 (highest median and smallest interquartile range – signifying greatest level of agreement) included: work ethic, passion for surgery, professionalism, ability to work in a team, sound judgment, ability to make decisions and assimilate information to formulate an opinion. Technical ability, although lower on the desired candidate attributes list, also met the criteria for inclusion (Table 13). Of the 10 open tasks and 15 VR tasks listed, 3 of each met the criteria for inclusion into selection, namely the one-handed tie, subcuticular interrupted suture, continuous suture and coordination, grasping and cutting. All other skills reached neutral consensus (Table 14).

4.4 Discussion

Selecting candidates for surgical training is a challenging and important responsibility that lacks empirical evidence to guide the process. Current selection in Canada consists of a structured review of a written application, followed by an in-person interview, with no formal assessment of incoming technical skill. With the current methods of selecting residents, the majority of PDs reported that 10-15% of graduating residents had difficulty reaching technical skill at the completion of surgical training. Despite this, technical skill is rarely incorporated into the selection process. In order to establish a collective opinion from PDs regarding desired candidate attributes and technical skills that are appropriate for addition to selection, a Delphi consensus was conducted. The results suggest that a number of different attributes are beneficial for success in GS residency. Furthermore, consensus was reached concerning a number of technical skills suitable for incorporation into selection. These included basic open tasks (one-handed tie, subcuticular interrupted suture and continuous suture) as well as basic laparoscopic skills (coordination, grasping and cutting).

The specific components included in current trainee screening and selection processes may explain the differences in the proportion of residents who had difficulty reaching competence in clinical knowledge (<5%), decision-making (5-10%) and technical skill (5-15%) by completion of surgical training, as reported by PDs. It appears that the least amount of trainees are reported to not reach competency in clinical knowledge, which may be attributed to the consideration of this domain within the existing education system and screening process. To qualify for application into residency training, candidates will have already repeatedly demonstrated academic excellence by retaining knowledge in order to score
in the top tier of students on written exams at the undergraduate level, as well as on The Medical College Admission Test (MCAT), which has been reported to predict training examination scores ("Admission requirements for Medical School at the University of Toronto," 2015; "Admission requirements University of British Columbia ", 2015; "Medical School Admission Requirement Dalhousie ", 2015; Ronai, Golmon, Shanks, Schafer, & Brunner, 1984). Furthermore, all candidates will have successfully completed 2-3 years of medical school, and thus will have demonstrated their academic potential in terms of learning and retaining clinical knowledge. However, the same screening processes are not available to assess decision-making and technical skill. In the current system, both of these domains have a more prominent focus in residency as compared to medical school, and therefore it is understandable that potential deficits are not uncovered until after admission. However, PDs report that fewer students have difficulty with decision-making as compared to technical skill. Perhaps this is also a result of the screening process currently in place, as decision-making is already incorporated in the interview component of the selection process at a majority (10 of 14) of institutions. Furthermore the interview score has been reported as a independent predictor of successfully completing GS training in a single study (Alterman, Jones, Heidel, Daley, & Goldman, 2011). In contrast, only 2 of 14 programs assess technical skill during selection. As a result, many trainees enter surgical programs with minimal or no screening in this domain. Given the notable proportion of trainees who PDs believed to have difficulty reaching competence, and that technical skill is an intrinsic part of surgical practice, it may be worth considering the incorporation of technical skill testing into selection.

A large number of candidate attributes were thought to be necessary for success in GS, however the current selection system poorly differentiates these qualities. In this study, 16 of 18 attributes listed by the PDs during the first round of the Delphi gained positive consensus during the second round. Currently, these attributes e.g. work ethic and professionalism, are evaluated at the time of interview with a subjective opinion rather than an objective measure. In an attempt to measure this qualitative area objectively, Bell et al. had 535 candidates across the Unites States complete the TriMetrix Personal Talent Report questionnaire: a questionnaire designed to assess behavioral motivators and student attributes. The authors then compared the candidate behavioral profiles’ to the final ranks lists at different institutions and found that individuals with diverse attributes can be similarly attractive candidates for surgical training (R. M. Bell, Fann, Morrison, & Lisk, 2011; Richard M. Bell, Fann, Morrison, & Lisk, 2012). Therefore, it seems that the current interview system is not able to separate candidates with specific attributes, but instead selects for a large range of different qualities. This may be partly due to
the subjective assessment method used in the current selection system to evaluate these qualities, coupled with the lack of empirical predictive evidence to inform which of these qualities actually increase the likelihood of success during residency training. Therefore to begin clarifying this somewhat vague area of selection, gaining expert consensus on the desired candidates attributes is an important first step. Further research is required to objectively assess these qualities and narrow the list with long-term supportive evidence.

Consensus was also gained on the technical skills that may be most appropriate for incorporation into the selection process, however, controversy remains as to how technical aptitude is best assessed in candidate trainees. To date, most studies have focused on the use of surrogate markers such as tests of dexterity and visual spatial ability to predict technical skill performance, based on the belief that these markers represent the building blocks required to excel in technical tasks ("Lafayette Instrument evaluation Dexterity Tests," 2015; D. Stefanidis, Korndorff Jr, et al., 2006). However, the results of studies to date have been disappointing. Although most surgeons would agree that high dexterity is of benefit in the operating room, scoring well on these tests has not correlated with technical performance (Hoffer & Hsu, 1990; Marlies P. Schijven et al., 2004; Schueneman et al., 1985; Schueneman et al., 1984; D. Stefanidis, Korndorff Jr, et al., 2006). Visual spatial ability has also been widely investigated in GS, especially since the advancements of in laparoscopic and endoscopic techniques. However, in both simulated laparoscopic and endoscopic tasks, the results are inconsistent (Ekstrom et al., 1976; A. G. Gallagher et al., 2003) (L. Enochsson et al., 2006; A. G. Gallagher et al., 2003; Groenier et al., 2014; E. Matt Ritter et al., 2006; D. Stefanidis, Korndorff Jr, et al., 2006). Therefore, although measuring aptitude through surrogate tests seemed to be a logical approach, it appears that technical performance is more complex than these tests are able to measure. Therefore, altering the approach to assess technical aptitude by incorporating simulation tasks that are directly transferable to the operating room is a worthwhile endeavor.

Assessing performance on simulated surgical tasks has been reported to differentiate trainees’ surgical aptitude. Grantcharov et al. plotted the individual learning curves of trainees performing basic VR simulator tasks, and noted that 8.1% of trainees lagged behind their peers and did not show any skill improvement (Grantcharov & Funch-Jensen, 2009). Similarly, Schijven et al. reported that 20% of residents did not reach proficiency on the laparoscopic clip and cut task after 30 trials (M. P. Schijven & Jakimowicz, 2004). These numbers are consistent with PD reports of 10-15% trainees having difficulty-reaching competency by the end of residency training. However, the evidence is limited to whether
performance on a technical skill during selection is predictive of future performance. Promising longitudinal data from the Otolaryngology program at the Mayo Clinic (Rochester, Minnesota) suggests that this may be the case. In this study, applicants to the program were asked to perform simple interrupted instrument ties under the microscope, with direct assessment from a faculty member using a GRS. This score was then incorporated into their total selection score (Carlson et al., 2010). Longitudinal follow up from this group suggested that incorporating the simulated simple interrupted microscopic suturing task at the time of selection was predictive of resident technical performance during training (Moore et al., 2014). Therefore, in order to assess whether GS can benefit from a similar process, introducing the open and laparoscopic skills outlined in this study should be considered.

We acknowledge two main limitations to our study. First, PDs from three programs did not respond to the questionnaire, and PDs from three additional programs contributed to the first round but not the second. Irrespective of this, the Delphi consensus methodology is robust to a moderate degree of attrition (Day & Bobeva, 2005). It has been reported that the expert panel will continue to produce useful results as long as between 7 and 10 members are retained (Day & Bobeva, 2005), which is lower than the number retained in our study. As a result, we believe that our results remain valuable and represent a national consensus on the topic of trainee selection into GS. Second, a large majority of responses concerning task selection fell within the category of neutral agreement. This may reflect considerable uncertainty among the expert panel as to whether this type of assessment is predictive of future performance. In the case of uncertainty, respondents are generally reluctant to express strong positive or negative opinions. Specifically in the case of the present study, PDs may be reluctant to express strong opinions concerning tasks that could potentially be used to eliminate trainees from entering GS when these tasks have not been tested for predictive validity. Nevertheless, the establishment of a national consensus around 6 skills that would be most useful to assess in candidates for GS training represents an important first step in undertaking further longitudinal study of their predictive value when incorporated into the overall selection process.

4.5 Conclusion

Selecting the best candidates for admission into residency is of upmost importance. Although PDs indicate that 5-15% of graduating residents had difficulty reaching competence in technical skill at the
completion of surgical training, assessment of this domain is rarely incorporated as part of the selection process. Consensus among PDs suggests that a number of both open (one-handed surgical tie, interrupted and running subcuticular sutures) and laparoscopic (coordination, grasping and cutting) skills are appropriate for inclusion. However, further assessment of whether these tasks are predictive of in-training performance is required before they can be recommended as a robust selection metric.
Chapter 5: Practice does not always make perfect - the need for selection curricula in modern surgical training

The content of this chapter has been submitted for review and has been peer reviewed by an expert in surgical education in simulation.

5.1 Abstract

**Background:** Emerging literature suggests that not all surgical trainees reach technical competence, although evidence to identify different learning patterns prior to entry into surgical training is limited. Therefore the purposes of the present study were to assess patterns of surgical skills acquisition among novice learners, to determine whether these were associated with baseline characteristics, and evaluate whether individuals’ learning patterns were consistent across a range of open and laparoscopic tasks of variable difficulty.

**Methods:** Sixty-five students were trained and completed a pre-determined, standardized training curriculum that included forty repetitions of the following laparoscopic and open technical tasks over a one month period. The tasks included peg transfer (PT), circle cutting (CC), intracorporeal knot tie (IKT), one-handed tie (HT) and simulated laparotomy closure (LC). A data mining technique, k-means clustering, was used to stratify the students into four performance learning clusters. Statistical analysis was used to compare the performance between the learning curves and further differentiate poor performers.

**Results:** Top performers (22-35% of participants) and high performers (32-42% of participants) reached proficiency in all tasks. Moderate performers (25-37% of participants) reached proficiency for all open tasks but not for all laparoscopic tasks. Low performers (8-15%) failed to reach proficiency four of five tasks including all laparoscopic tasks (PT task 7.8%; CC task 9.4%; and IS task 15.6%). For laparoscopic tasks, participants in higher performance clusters demonstrated innate abilities that conferred significantly better initial performance, as well as a significant sustained performance advantage across
all repetitions. In contrast, the majority of moderate or low performers remained in these categories across all the skill types and complexities studied. Low performers’ learning curves were widely variable, and lacked evidence of progression towards a plateau phase.

**Conclusion:** Whilst most students will reach proficiency with continued practice and mentorship across a range of surgical tasks, low performing trainees failed to reach proficiency with laparoscopic tasks, demonstrating highly variable performance scores across the entire learning curve. Given the increasing use of laparoscopic techniques in surgical practice, screening potential candidates to identify the lowest tier performers may benefit both students and their training programs.

### 5.2 Background

A common paradigm in surgical education is that all surgical trainees will reach technical competence with continued practice and adequate mentorship. (K.A. Ericsson, Krampe, & Tesch-Römer, 1993; G., 2008; Gladwell, 2008) However, emerging evidence from the minimally invasive simulation literature contests this belief and supports the notion that trainees acquire technical skills at variable rates, with a subset of students unable to reach competence (Alvand, Auplish, Khan, Gill, & Rees, 2011; Cushchieri, 2003; Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). Recent studies suggest that 5% - 17% of trainees have an innate technical ability that allows them to rapidly acquire skills, achieving competence with minimal practice or effort (top/high performers) (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). These students quickly and effectively learn the technical components of the operation, allowing them to focus their attention on other competencies such as surgical judgment and non-technical performance. In contrast, most trainees (63-70%) are moderate performers (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). They improve with diligent practice, ultimately reaching a level of technical proficiency that is acceptable and safe. However, studies have also identified a smaller subgroup of trainees (8-20%) who struggle to learn technical skills and fail to reach competence even with continued practice (low performers).(Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004) Given that technical skill is a requirement for a successful surgical career, identifying these individuals early may benefit both prospective trainees and surgical training programs.

The selection process for surgical trainees in North America does not routinely include screening for
technical ability. This may be because of controversy as to whether low performers can be accurately identified at this early stage of training, and uncertainty whether they are truly disadvantaged in the clinical environment (Louridas, Szasz, de Montbrun, Harris, & Grantcharov, 206). To date, attempts at stratifying trainees into high and low performers has been done by assessing individual learning curves for basic laparoscopic surgical tasks, or by assessing performance on a single simulated operation (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). However, these may not realistically model the diversity of the surgical training environment.

In residency, surgical trainees generally progress from basic to more advanced technical skills. It is unknown whether trainees, classified as poor performers on basic tasks, struggle across tasks of variable difficulty, or whether they are able to overcome these shortcomings given diverse technical experiences. Furthermore, there are few reported data concerning variability in performance for open surgical skills. Given that many surgical disciplines combine minimally invasive and open techniques in their clinical practice, an understanding of the learning progression for both types of skills over the continuum of skill progression (from basic to more advanced) is needed before it can be determined whether initially low performing individuals will remain disadvantaged throughout their training.

Thus, the overall purpose of this study was to assess the differences in technical performance by novice trainees across a range of surgical tasks. Specifically, the goals were to: 1) quantify different learning patterns among trainees for both basic and more advanced laparoscopic and open skills; 2) assess whether background characteristics or experiences explained potential differences in performance; 3) determine whether trainees stay within their learning curve patterns across simulated tasks of varying difficulty (basic and advanced) and type (minimally invasive and open); and 4) identify the subset of trainees who consistently fail to reach proficiency on simulated tasks, and determine the features of their learning curves that separate them from their peers.

5.3 Methods

5.3.1 Participants

Medical students completing their first or second year at the University of Toronto were eligible to participate in the study. All students willing to commit to one month of technical skills training were deemed eligible for inclusion. To ensure that participants were at the beginning of their technical skills
learning curve, the following exclusion criteria were applied: 1) previous involvement in a dedicated technical skills training program, or 2) participation in more than two suturing workshops offered during the pre-clerkship medical school curriculum. Ethics review board approval was obtained through the University of Toronto, and informed consent was obtained from each participant prior to the start of the study.

5.3.2 Sample Size

An \textit{a priori} sample size calculation was performed by modifying the equation commonly used for studies that estimate the proportion of a population with specified precision for a binary outcome (Merril, 2015). The primary outcome measure for this study was the proportion of participants demonstrating improvement in technical skill following training. Consequently, participants could be found to be either responders or non-responders for the primary outcome measure. The literature suggests that 8-20\% of surgical trainees are non-responders. Therefore, the proportion of non-responders in the overall population (p) was assumed to be 15\% for the purposes of the sample size calculation (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004; Sir Alfred, 2003). To ensure a 95\% likelihood that the study sample included at least 4 non-responders (ie. assuming a sample size of n and precision of d, that n(p − d) = 4), the desired precision was defined as d = p − 4/n. Using the equation for estimating population proportions with a given precision

\[ n = \frac{z^2 p(1-p)}{d^2} \]

where n = sample size, \( z = 1.96 \) (for 95\% confidence interval), \( p = 0.15 \), and \( d = p-4/n \), a desired sample size of 64 participants was calculated.

5.3.3 Demographic questionnaire

At the start of the study each participant completed a questionnaire assessing a range of demographic characteristics and previous non-surgical experiences that may potentially be associated with technical skill performance. Demographic characteristics included age, gender and handedness, whereas non-surgical experiences included interest in surgery, self-perceived technical ability, and involvement in music, sports and video games.
5.3.4 Training curriculum

5.3.4.1 Structure of skills curriculum

Students completed a one month distributed practice training program designed to teach both open and laparoscopic skills. The program comprised a minimum of fourteen training sessions scheduled 2-3 times per week. The first two weeks were dedicated to basic skills and the last two weeks to more advanced skills. At the beginning of each two week block, the students first watched a detailed instructional video for each task, and then completed a baseline performance test. Following the baseline test, each student then completed four more repetitions of each task. At each subsequent session, seven repetitions of each task were performed, for a total of 40 repetitions per task.

5.3.4.2 Instructors and feedback

The student to instructor ratio was maintained at 4:1 for each training session consistent with recommendations in the literature.(Dubrowski & MacRae, 2006) To minimize the impact of the differences between trainers on the students’ technical performance, only three different instructors taught the skills over the study period, and instructors gave standardized feedback after each repetition using a task-specific checklist.

5.3.4.3 Data collection

Each repetition was video recorded to allow for subsequent blinded assessment. For the laparoscopic tasks, the inside view of the training box was captured by the laparoscopic camera. For the open tasks, a portable video recording device was used to capture only the students’ hands and simulated operative field. For the purposes of assessment and analysis, a paper containing a unique code was placed within the field of view of the camera to allow each recorded attempt to be tied to an individual participant and repetition number in a de-identified manner.
5.3.4.4 Technical skills

The peg transfer (PT) was used as the basic laparoscopic task, and the one-handed tie (HT) was used as the basic open task. The peg transfer was performed as outlined in the Fundamentals of Laparoscopic Surgery (FLS) manual."FLS Manual Skills Written Instructions and Performance Guidelines," 2014) The HT was performed using a silk tie suture passed through a Penrose drain fastened with Velcro to an underlying board.

The circle cut (CC) and intracorporeal knot tie (IKT) were used as the advanced laparoscopic tasks, and a simulated open laparotomy closure (LS) was used as the advanced open task. The advanced laparoscopic tasks were performed as outlined in the FLS manual."FLS Manual Skills Written Instructions and Performance Guidelines," 2014) The open LC was simulated by fastening a felt sheet to an underlying hollow frame with a 15cm incision down the middle of the felt. A 0-monofilament suture was used to close the incision. This model was modified from the ACS/APDS surgical skills curriculum(Cofer, Lee, Pritts, & Harvey, 2015).

5.3.4.5 Assessment of technical skills

For the HT (basic open task), students were asked to tie as many square knots as possible in one minute. A modified Objective Structure of Technical Skills (mOSATS) scale was used to score each repetition (Martin et al., 1997). This modified scale excluded the ‘use of assistance’ and ‘knowledge of instruments’ categories included in the original OSATS assessment tool, as they were not applicable to this task(Martin et al., 1997). During the advanced open task sessions, students were asked to perform a simulated LC in 300 seconds. Each repetition was scored using a mOSATS.

The open task videos were assessed at the completion of the study, to allow for all the videos to be rearranged and assessed in a random order. A total of three expert blinded raters assessed the open tasks. Before rating was begun, these individuals were trained on the mOSATS instruments and their ratings were calibrated to achieve excellent inter-rater reliability, defined as an intraclass correlation (ICC) coefficient of >0.75 (Downing, 2004). To ensure that calibration was maintained over the three-month rating period, the raters’ ICC was re-calculated every two weeks. Excellent reliability was maintained, with ICCs ranging from 0.78 to 0.92 over the rating period.
All laparoscopic tasks were scored in real-time to remain consistent with the validated methodology of the FLS scoring system. Time and error scores were collected for each task as outlined in the FLS manual ("FLS Manual Skills Written Instructions and Performance Guidelines," 2014). The videos were used to check the recorded time scores, and to fill in any unrecorded values to shield against missing data. Given that the FLS scoring of time and error can be objectively measured, blinded rating in a random order, as done with the open tasks, was not necessary.

5.3.4.6 Performance goals

Students were trained to 40 repetitions for each basic and advanced task, with the goal of reaching proficiency as quickly as possible. Once proficiency was reached, emphasis was placed on maintenance of these scores during subsequent repetitions.

The scores published by Ritter et al. were used as the proficiency cut offs for the laparoscopic tasks (E. M. Ritter & Scott, 2007). The published cutoffs for the CC and IKT tasks were defined by Ritter et al. as two standard deviations below (slower than) the expert mean, while the cutoff for PT was reported as the mean only. Consequently, the cutoff for the PT score in the present study was defined as two standard deviations below (slower than) the mean, to keep it consistent with the other FLS tasks. The specific proficiency scores used were: 54.8 seconds for the PT; 98.0 seconds for the CC, and 113.0 seconds for the IKT. These time scores were then normalized for analysis to 103.4, 72.1, and 93.8 seconds respectively, as per the validated FLS scoring system methodology (Vassiliou et al., 2006).

No previously published proficiency scores were identified for the open tasks. Consequently, the methodology described by Ritter et al. for setting FLS task performance targets was used to establish proficiency scores for the open tasks (E. M. Ritter & Scott, 2007). Two expert surgeons from the study institution performed each basic and advanced task five consecutive times. Each repetition was evaluated using the associated mOSATS instrument by a single blinded rater. The completion times for all 10 repetitions were used to calculate the mean and standard deviation for each task, with two standard deviations below (slower than) the mean set as the proficiency score for each task (Table 15). Per the scoring methodology, higher scores represented better performance for all tasks.
Table 15: Proficiency scores for open tasks

<table>
<thead>
<tr>
<th>Repetition</th>
<th>One handed tie mOSATS</th>
<th>Laparotomy closure mOSATS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Expert 1</td>
<td>Expert 2</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>25</td>
</tr>
</tbody>
</table>

Mean | 23 | 23 | 27.2 | 25.8 |
SD   | 1.6| 1.6| 2.2  | 2.2  |

Group mean | 23 | 26.5 |
Group SD    | 1.5| 2.2  |
Proficiency score | 20.0 | 22.2 |

SD: standard deviation; mOSATS: modified objective structure of technical skills; Proficiency score = Group mean - 2(SD)

5.3.5 Data analysis and statistical methods

A data mining technique, k-means clustering, was used to stratify the trainees into four categories. Data mining is a computational process of finding patterns in various types of data using a combination of methods from artificial intelligence and machine learning (Alpaydin, 2014). K-means clustering is a technique used to partition \( n \) observations into \( k \) clusters, where observations within a cluster display greater similarity to one another than to those in other clusters (Tan, Steinbach, & Kumar, 2006). The surgical education literature has previously separated trainees into four performance categories therefore \( k=4 \), was selected for this analysis (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). While the number of clusters with this analytic approach is defined \( a \) priori, the number of
observations in each cluster varies and is determined by the relationships between observations. The cluster groupings are determined through a process of initial and updated assignment of observations to clusters, with the goal of achieving the lowest sum of squares of the Euclidean distance from the cluster mean. This analysis is entirely objective, and is able to account for variation in each individual’s learning process before finalizing placement into the appropriate cluster for that task. K-means clustering analyses were conducted using R statistical software (version 3.2.2; R Foundation for Statistical Computing, Vienna Austria).

To ensure the identified clusters accurately represented technical performance levels, each cluster’s mean learning curve was graphed on a traditional x and y-axis to evaluate whether each group approached, reached, or surpassed the set proficiency threshold for each task.

Chi-squared analysis was then used to determine whether the same trainees fell within the same performance clusters between tasks, and to assess whether one or more demographic and background characteristics had a significant relationship with technical skill.

Consistency of each participant’s performance at the conclusion of training was quantified by examining the performance scores for their last 10 repetitions for each task, excluding the highest and lowest score (to minimize skewing due to outlier values), and calculating the performance range of the remaining eight repetitions (highest score minus lowest score).

All statistical analyses were performed using SPSS version 23 (IBM SPSS Statistics; IBM Corporation, Armonk, NY).

5.4 Results

5.4.1 Participants

Sixty-five medical students completed the study. Sixty-six students were enrolled, but one dropped out due to scheduling conflicts. The mean age of the participants was 24 years (SD = 2.2), and there were 29 males and 36 females.
Figure 11: Learning curves for clusters 1 (top performers) to 4 (low performers) on the laparoscopic (a) peg transfer, (b) circle cut, and (c) intracorporeal knot tie tasks, ordered from basic to advanced. Proficiency scores are demarcated with a dashed line.

5.4.2 Learning curves patterns for laparoscopic and open skills

Based on longitudinal performance over all repetitions, participants were effectively separated into four distinct clusters, corresponding to four learning curves, for each basic and advanced task. Cluster assignments were determined independently for each task, without consideration of their performance across different skills. For each task, top performers were labeled as cluster 1, high performers as cluster 2, moderate performers as cluster 3 and low performers as cluster 4 (Figure 11 and Figure 12).
Figure 12: Learning curves for clusters 1 (high performers) to 4 (low performers) on the open (a) one handed tie and (b) laparotomy closure tasks, ordered from basic to advanced.
Learning curves for cluster 1 (22 to 35% of participants) were characterized by rapid achievement of proficiency for all open and laparoscopic tasks, which was then reliably maintained on subsequent repetitions. Learning curves for participants in cluster 2 (32 to 42% of participants) were characterized by quick learning, eventually meeting the performance of their cluster 1 peers and reaching proficiency for all tasks over the training period. Cluster 3 learning curves (25 to 37% of participants) were characterized by slower learning, with proficiency attained for the open tasks but not for all the laparoscopic tasks. Specifically, cluster 3 performance remained just below the proficiency threshold for both PT and IKT tasks. Cluster 4 learning curves (8 to 15% of participants) were characterized by failure to reach proficiency in four of five tasks, including all laparoscopic tasks (Table 16).
### Table 16: Comparing clusters 1-4: start points, end points and repetitions to proficiency

<table>
<thead>
<tr>
<th>Simulated surgical task</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Comparing clusters 1-4 Kruskal-Wallis H Test</th>
<th>p-value</th>
<th>The clusters that are significantly different</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Laparoscopic tasks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peg transfer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of students</td>
<td>23</td>
<td>21</td>
<td>16</td>
<td>5</td>
<td></td>
<td>*&lt;0.001</td>
<td>all 4 clusters</td>
</tr>
<tr>
<td>Start point score (points) mean (range)</td>
<td>75 (60-89)</td>
<td>46 (17-65)</td>
<td>29 (27-56)</td>
<td>17 (0-40)</td>
<td>49.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End point score (points) mean (range)</td>
<td>108 (103-113)</td>
<td>105 (99-109)</td>
<td>95 (85-103)</td>
<td>90 (83-96)</td>
<td>47.78</td>
<td>*&lt;0.001</td>
<td>all 4 clusters</td>
</tr>
<tr>
<td>Proficiency reached repetition (range)</td>
<td>16 (8-26)</td>
<td>26 (20-33)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Circle cut</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of students</td>
<td>15</td>
<td>25</td>
<td>19</td>
<td>6</td>
<td></td>
<td>*&lt;0.001</td>
<td>cluster 1 and 2, cluster 2 and 3</td>
</tr>
<tr>
<td>Start point score (points) mean (range)</td>
<td>54 (34-73)</td>
<td>43 (23-57)</td>
<td>22 (0-36)</td>
<td>25 (17-39)</td>
<td>44.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End point score (points) mean (range)</td>
<td>91 (85-97)</td>
<td>88 (72-107)</td>
<td>82 (69-87)</td>
<td>71 (57-82)</td>
<td>33.82</td>
<td>*&lt;0.001</td>
<td>all 4 clusters</td>
</tr>
<tr>
<td>Proficiency reached repetition (range)</td>
<td>5 (3-6)</td>
<td>8 (4-11)</td>
<td>16 (8-24)</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intracorporeal knot tie</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of students</td>
<td>14</td>
<td>27</td>
<td>14</td>
<td>10</td>
<td></td>
<td>*&lt;0.001</td>
<td>all 4 clusters</td>
</tr>
<tr>
<td>Start point score (points) mean (range)</td>
<td>43 (11-70)</td>
<td>25 (1-60)</td>
<td>13 (0-36)</td>
<td>1 (0-2)</td>
<td>39.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 1</td>
<td>Cluster 2</td>
<td>Cluster 3</td>
<td>Cluster 4</td>
<td>Significance</td>
<td>Clusters</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td><strong>End point score (points)</strong></td>
<td>95 (87-102)</td>
<td>91 (77-100)</td>
<td>89 (64-100)</td>
<td>78 (55-93)</td>
<td>17.28</td>
<td>*0.001</td>
<td>all 4 clusters</td>
</tr>
<tr>
<td>Mean (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proficiency reached</strong></td>
<td>29 (17-35)</td>
<td>35 (25-40)</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Open technical skills**

**One handed tie**

**(mOSATS)**

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Significance</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of students</strong></td>
<td>20</td>
<td>23</td>
<td>16</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Start point score (points)</strong></td>
<td>9 (6-14)</td>
<td>8 (7-12)</td>
<td>7 (7-8)</td>
<td>7 (7-9)</td>
<td>7.09</td>
<td>0.069</td>
</tr>
<tr>
<td>Mean (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>End point score (points)</strong></td>
<td>23 (19-25)</td>
<td>22 (19-24)</td>
<td>20 (18-24)</td>
<td>17 (14-20)</td>
<td>28.56</td>
<td>*&lt;0.001</td>
</tr>
<tr>
<td>Mean (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proficiency reached</strong></td>
<td>17 (13-28)</td>
<td>24 (21-30)</td>
<td>34 (28-38)</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Laparotomy closure**

**(mOSATS)**

<table>
<thead>
<tr>
<th></th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
<th>Significance</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of students</strong></td>
<td>14</td>
<td>26</td>
<td>20</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Start point score (points)</strong></td>
<td>13 (9-21)</td>
<td>12 (9-17)</td>
<td>12 (8-17)</td>
<td>11 (8-14)</td>
<td>1.96</td>
<td>0.581</td>
</tr>
<tr>
<td>Mean (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>End point score (points)</strong></td>
<td>27 (21-30)</td>
<td>24 (19-29)</td>
<td>24 (18-29)</td>
<td>23 (20-27)</td>
<td>9.00</td>
<td>*0.029</td>
</tr>
<tr>
<td>Mean (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Proficiency reached</strong></td>
<td>16 (3-29)</td>
<td>27 (13-39)</td>
<td>26 (15-38)</td>
<td>36 (34-38)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition (range)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

mOSATS: modified Objective Structure of Technical Skills; *significance p<0.05; start point - average of first 3 repetitions; end point - average of last 3 repetitions.
Innate technical ability seemed to contribute significantly to laparoscopic performance, but not to open skills. In all three laparoscopic tasks, students in higher clusters had significantly better mean performance at the start of the learning curve as compared to peers in the next lower group (mean initial performance for clusters 1 through 4 on the PT task of 75, 46, 29 and 4 points respectively, \(p<0.001\); on the CC task of 54, 43, 22 and 25 points respectively, \(p<0.001\); on the IKT task of 43, 25, 13 and 1 points respectively, \(p<0.001\)). The only exceptions were clusters 3 and 4 for the CC task, which were characterized by similar performance (Table 16). For the open tasks, starting performance scores were similar across all clusters (mean initial performance for clusters 1 through 4 on the HT task of 9, 8, 7 and 7 points respectively, \(p=0.069\); on the LC task of 13, 12, 12 and 11 points respectively, \(p=0.581\)).

The innate performance advantage for laparoscopic tasks seen in higher performing clusters persisted over the duration of the study. The associated mean laparoscopic learning curves remained distinctly higher in comparison to those of lower performing clusters throughout the 40 repetitions for all three tasks (Figure 11). This resulted in significantly different mean training endpoints between clusters (mean final performance for clusters 1 through 4 on PT task of 108, 105, 95 and 90 points respectively, \(p<0.001\); on CC task of: 91, 88, 82 and 71 points respectively, \(p<0.001\); on IKT task of 95, 91, 89 and 78 points respectively, \(p=0.001\)) (Table 16).

5.4.3 Associations between background characteristics and performance clusters

Only two significant associations were identified between participant demographic characteristics, non-surgical experiences, and performance clusters across all five technical tasks (Table 17). Both significant associations were noted with respect to the peg transfer. Specifically, males were more likely to be in higher performance clusters compared to females (cluster 1 had 65% males and 35% females as compared to 45% vs. 55% overall, \(p=0.039\)), and students who played video games were more likely to be in a higher performance cluster compared to students who did not play video games (cluster 1 had 73% video gamers and 27% non-gamers as compared to 52% and 48% overall, \(p=0.021\)).
Table 17: Demographics and non-surgical experiences and their association with performance clusters

<table>
<thead>
<tr>
<th>No. of participants n=65</th>
<th>Laparoscopic Peg Transfer Task</th>
<th>Laparoscopic Circle Cut Tasks</th>
<th>Laparoscopic Knot Tie Task</th>
<th>Open One Handed Tie Task</th>
<th>Open Laparotomy Closure Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Performance cluster</td>
<td>p-value</td>
<td>Performance cluster</td>
<td>p-value</td>
<td>Performance cluster</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
<td>1 2 3 4</td>
</tr>
<tr>
<td>Medical school year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>49 (75)</td>
<td></td>
<td>17 (26)</td>
<td></td>
<td>17 (26)</td>
</tr>
<tr>
<td>Two</td>
<td>16 (25)</td>
<td></td>
<td>13 (20)</td>
<td></td>
<td>15 (23)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>29 (45)</td>
<td></td>
<td>15 (23)</td>
<td></td>
<td>7 (11)</td>
</tr>
<tr>
<td>Female</td>
<td>36 (55)</td>
<td></td>
<td>9 (14)</td>
<td></td>
<td>8 (12)</td>
</tr>
<tr>
<td>Handedness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>59 (92)</td>
<td></td>
<td>20 (31)</td>
<td></td>
<td>14 (22)</td>
</tr>
<tr>
<td>Left</td>
<td>5 (8)</td>
<td></td>
<td>3 (5)</td>
<td></td>
<td>1 (2)</td>
</tr>
<tr>
<td>Played a musical instrument</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>44 (69)</td>
<td></td>
<td>16 (25)</td>
<td></td>
<td>12 (19)</td>
</tr>
<tr>
<td>No</td>
<td>20 (31)</td>
<td></td>
<td>6 (9)</td>
<td></td>
<td>3 (5)</td>
</tr>
<tr>
<td>Involvement in team sports</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>46 (71)</td>
<td></td>
<td>16 (25)</td>
<td></td>
<td>12 (19)</td>
</tr>
<tr>
<td>No</td>
<td>19 (29)</td>
<td></td>
<td>11 (17)</td>
<td></td>
<td>9 (14)</td>
</tr>
<tr>
<td>Work well with my hands</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Agree</td>
<td>45 (70)</td>
<td></td>
<td>18 (28)</td>
<td></td>
<td>13 (20)</td>
</tr>
<tr>
<td>Neutral</td>
<td>19 (30)</td>
<td></td>
<td>7 (11)</td>
<td></td>
<td>5 (8)</td>
</tr>
<tr>
<td>Missing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Games</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>33 (52)</td>
<td></td>
<td>16 (25)</td>
<td></td>
<td>9 (14)</td>
</tr>
<tr>
<td>No</td>
<td>31 (48)</td>
<td></td>
<td>9 (14)</td>
<td></td>
<td>6 (9)</td>
</tr>
<tr>
<td>Surgery primary interest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25 (38)</td>
<td></td>
<td>9 (14)</td>
<td></td>
<td>6 (9)</td>
</tr>
<tr>
<td>No</td>
<td>40 (62)</td>
<td></td>
<td>14 (22)</td>
<td></td>
<td>9 (14)</td>
</tr>
</tbody>
</table>

Chi-square Test was used throughout the table; *significant p-value <0.05

Chi-square Test was used throughout the table; *significant p-value <0.05
5.4.4 Individual performance differences between tasks

A significant association was seen between individuals’ cluster assignments across tasks (basic vs advanced) within a given skill type (laparoscopic vs open), indicating that individuals were likely to remain within the same or similar performance cluster across multiple skills of the same type. However, this phenomenon was not demonstrated across the two skill types (e.g. low performers in laparoscopic tasks were not necessarily low performers in open tasks) (Table 18).

Within a given skill type, individuals who were top performers (cluster 1) in the basic tasks stayed in the upper tier (either cluster 1 or 2) for all basic and advanced tasks. Cluster 2 individuals maintained their grouping or moved up or down a cluster depending on the task. However, poor performers, stayed in the lower tier (either cluster 3 or 4) for all tasks. Only one individual did not follow this trend, and jumped from one extreme to the other (cluster 4 to cluster 1) between tasks (Table 18).
Table 18: Consistency of performance clusters for laparoscopic and open technical skill

<table>
<thead>
<tr>
<th>Participants that stayed within:</th>
<th>Clusters 1-2</th>
<th>Cluster 2-3</th>
<th>Clusters 3-4</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consistency of task clusters 1 - 4, between tasks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laparoscopic peg transfer</td>
<td>Laparoscopic circle cut</td>
<td>95%</td>
<td>66%</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>Laparoscopic knot tie</td>
<td>71%</td>
<td>66%</td>
<td>38%</td>
</tr>
<tr>
<td>Laparoscopic circle cut</td>
<td>Laparoscopic knot tie</td>
<td>63%</td>
<td>77%</td>
<td>42%</td>
</tr>
<tr>
<td>Open hand tie</td>
<td>Open Laparotomy closure</td>
<td>73%</td>
<td>74%</td>
<td>44%</td>
</tr>
<tr>
<td><strong>Consistency of task clusters 1 - 4, across surgical techniques</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Open skills</strong></td>
<td><strong>Laparoscopic skills</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hand ties</td>
<td>Peg transfer</td>
<td>60%</td>
<td>65%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>Circle cut</td>
<td>18%</td>
<td>64%</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>Knot tie</td>
<td>62%</td>
<td>59%</td>
<td>24%</td>
</tr>
<tr>
<td>Laparotomy closure</td>
<td>Peg transfer</td>
<td>61%</td>
<td>73%</td>
<td>38%</td>
</tr>
<tr>
<td></td>
<td>Circle cut</td>
<td>65%</td>
<td>70%</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>Knot tie</td>
<td>63%</td>
<td>68%</td>
<td>42%</td>
</tr>
</tbody>
</table>
5.4.5 Concerning learning curve features for individuals unable to reach proficiency in laparoscopic tasks

For laparoscopic tasks, cluster 3 individuals characterized by slower learning did not reach proficiency in two of three laparoscopic tasks, while cluster 4 individuals did not reach proficiency in any of the laparoscopic tasks. When comparing the two groups, the biggest difference was performance stability between repetitions. Cluster 3 hovered below the proficiency threshold but demonstrated an encouraging trend towards a plateau phase, indicative of more stable performance (mean range in scores across last 10 repetitions PT=10.36; CC= 8.9 and IKT 13.8). In contrast, cluster 4 participants demonstrated highly variable scores between repetitions, with no convincing trend towards a true plateau phase, indicative of more unstable performance (mean range in scores of last 10 repetitions: PT=13.2; CC= 20.0 and IKT=23.77).

The wide variability in cluster 4 participants’ performance is a noteworthy and concerning feature of their learning curves that may translate into a negative impact in the real operating room. Figure 13 represents the IKT learning curve for representative individuals from clusters 3 and 4. In the last 10 repetitions per knot, the cluster 3 individual’s normalized score was between 103 and 154 points (equivalent to a performance time of 1.43 to 2.34 minutes), whereas the cluster 4 individual scored between 174 and 456 points (performance time of 2.54 to 7.36 minutes). The performance times seen in cluster 4 (frequently >4 minutes per knot) are likely to be disruptive to the progression and flow of a laparoscopic operation. Therefore, although individuals in both clusters do not stabilize above the proficiency threshold, cluster 4’s inconsistent non-proficient performance would likely be more problematic in the real operating room.
Figure 13: Learning curves for a representative individual from cluster 3 (moderate performer) and from cluster 4 (low performer) on the laparoscopic intracorporeal knot tie, demonstrating variability in performance times and associated normalized scores.

5.5 Discussion

This study was undertaken as recent literature suggests that not all surgical trainees may be able to reach technical proficiency (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004). However, there has been no implementation of selection curricula in clinical practice due to lack of sufficient evidence on the learning patterns of trainees across a range of both laparoscopic and open techniques of variable difficulty. Assessing a variety of tasks as was done in this study, may be a more realistic reflection of the fundamental skill sets required for surgical training. Therefore, trainees were stratified into performance clusters for basic and advanced, laparoscopic and open tasks. Performance clusters were consistent between skill types (basic to advanced) but not across surgical techniques (open versus laparoscopic). Our results suggest that laparoscopic technical skills were more challenging to acquire and presented a greater barrier to reaching proficiency as compared to open skills. Low performing trainees in
the laparoscopic tasks demonstrated weak innate ability with a sustained disadvantage throughout their learning curves resulting in an inability to reach proficiency within the extensive training window. Furthermore, laparoscopic low performers demonstrated marked inconsistencies in their performance throughout the training period.

K-means clustering, which is a novel data mining technique within the surgical education literature, but well established within the field of data science, was used in the present study instead of previously described approaches that characterize individual learning curves such as the learning curve – cumulative sum analysis (LC-CUSUM) or curve fitting (Biau, Williams, Schlup, Nizard, & Porcher, 2008; M. P. Schijven & Jakimowicz, 2004). LC-CUSUM is applicable to the analysis of dichotomous variables (Biau et al., 2008), and allows researchers to identify trainees who reach proficiency with performance variability below a pre-defined threshold. However, little additional information is gained to further characterize the performance of trainees that do not reach proficiency. Curve fitting endeavors to address this issue by classifying learners’ individual learning curves by fitting an exponential, linear, logarithmic, or S curve to individual learners’ change in performance over time, allowing for stratification based on similar curve types (M. P. Schijven & Jakimowicz, 2004). However, because many trainees’ learning curves demonstrate considerable residual variability, the discriminatory value of statistics used to quantify the goodness-of-fit of learning curves to different curve types is poor (Feldman, Cao, Andalib, Fraser, & Fried, 2009; M. P. Schijven & Jakimowicz, 2004). In contrast, k-means clustering allows the observations to be objectively separated based on patterns within the data, rather than relying on force fitting of data to one of a limited number of pre-defined mathematical curves. Furthermore, the analysis is not restricted to a dichotomous interpretation (ie. did or did not meet proficiency), and detailed comparisons between clusters can be performed to further quantify performance differences between the groups.

Many surgeons acknowledge that some trainees have inherent talent and thus demonstrate quick acquisition of technical skill (Anthony G. Gallagher et al., 2009). However, conceding that a subset of trainees will likely never reach proficiency is contentious. Perhaps this controversy exists because consistent inability to reach proficiency may only be true for a subset of skills (i.e. laparoscopic and not open). On examining the learning curves of the different tasks, all students reached proficiency for the advanced open task, albeit at different rates. In contrast, practice did
not close the performance gap between clusters for the laparoscopic tasks even over 40 repetitions. The current study purposefully lengthened the training period beyond the 10 to 30 repetitions previously reported in the literature (A. Alvand, S. Auplish, T. Khan, et al., 2011; Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004), in an attempt to determine whether more practice would allow low performers to catch up. Despite doing so, the performance clusters remained disparate throughout the learning curve for laparoscopic tasks. Students with low innate ability remained at a continued disadvantage throughout the 40 repetitions. Furthermore, the proportion of low performers did not change substantially when compared to previously reported studies, despite the added practice. For basic tasks, Grantcharov et al. reported 8.1% of trainees to be low performers, compared to 7.6% in the current study (Grantcharov & Funch-Jensen, 2009). For more difficult tasks, Schijven et al. reported a 20% rate of low performers on virtual reality laparoscopic cholecystectomy, (M. P. Schijven & Jakimowicz, 2004) compared to 15.4% for intracorporeal suturing in the current study. These findings provide further evidence that low performers persist across laparoscopic tasks. They exhibit low innate ability, with a lasting effect as they continue to perform below the proficiency threshold despite lengthening the practice window.

The possibility of screening for individuals who are low performers in technical skills prior to entry into surgical training is controversial. The results of the present study suggest that screening may be of limited benefit for surgical specialties with predominately open operations, given that all participants were able to reach proficiency on at least one of these tasks. However, in specialties where an increasing number of laparoscopic procedures are becoming the standard of care (Soper, Stockmann, Dunnegan, & Ashley, 1992; Taguchi et al., 2016), screening for performance on these technical skills may be beneficial. In the present study, the large majority of low performers continued to struggle throughout the curriculum across a large number of repetitions and multiple laparoscopic tasks. Furthermore, the inconsistency in performance between repetitions in this subgroup was substantial, despite completing each repetition in a controlled simulated environment where the setup, equipment and simulated task were constant. It is reasonable to believe that this poor and inconsistent performance would be exacerbated in the real operating room, where patient factors (e.g. body habitus, tissue integrity, patient anatomy), technical setup (e.g. trocar placement, positioning and camera view), and trainee feedback may vary substantially. Thus, although controversial, it may be reasonable to consider
screening candidates entering surgical specialties with considerable laparoscopic procedure volumes to identify those individuals who cluster into the lowest performance category. These individuals are at high risk of ongoing variable performance below an acceptable level of proficiency, which could have negative implications for both the training program and the trainee.

Using surrogate markers instead of simulated tasks to screen for technical skill has also been extensively investigated, but has proven to be largely unreliable (Louridas et al., 2006). Currently all students in North America, applying to surgical training, are required to submit a curriculum vitae outlining their academic and non-academic achievements (Canadian Resident Matching Service, 2014; NRMP, 2016). In the current study, very few of these background characteristics demonstrated significant associations with either technical skill acquisition or the ability to reach proficiency. Interestingly, students who reported that surgery was their first choice of future career specialty had no added performance advantage as compared to their peers who expressed a preference for non-surgical disciplines. Similarly, students who reported “working well with their hands” had no advantage over their peers. This data adds to the already existing evidence that self-perceived technical ability does not correlate with objective technical assessments (Tangchitnob et al., 2011). Furthermore, it has been demonstrated that medical students intending to pursue a surgical career have no greater innate motor dexterity when compared to those with a preference for a non-surgical field (J.Y. Lee et al., 2012; Van Hove et al., 2008). Consequently, surgical programs should not rely on student self-selection as a screening tool (Panait et al., 2011). Of the other background data, playing video games was significantly associated with technical skill (Schueneman et al., 1985; Van Hove et al., 2008). However, the technical advantage associated with playing video games was lost after the basic laparoscopic PT task, which is consistent with previous reports by Paschold et al. and Dimitriou et al. that demonstrated that any advantage associated with video game playing was lost after a period of practice (Dimitriou et al., 2009; Paschold et al., 2011). Therefore, although surrogate markers are an attractive screening tool due to their feasibility in the current selection process, such markers are not predictive of technical skill acquisition and thus should not be used in this manner.

We acknowledge several limitations to the present work. First, the study population consisted of pre-clerkship medical students, not only medical students applying for surgical residency. This population was chosen because within the Canadian medical system, only medical students in the
pre-clinical years of study are reliably at the start of their technical skill learning curves, a requirement given the purposes of the present work. In addition, pre-clerkship medical students could be trained solely in the simulation laboratory, capturing each repetition of the study tasks as they progressed along their learning curves without the risk of confounding as a consequence of undocumented practice in the operating room. Such controlled circumstances would not have been possible if studying final year medical students. Nevertheless, the authors believe these results are both useful and relevant to medical students applying for surgical training. As previously described, candidates for surgical training do not self-select based on technical skill. Consequently, surgical programs are likely interviewing and accepting trainees representing all four clusters identified in the present study. Furthermore, since 40 repetitions per task were used, the skill acquired and reflected in the proficiency plateaus by the study participants would likely surpass the skill (and subsequent plateaus) of medical students entering the training program. Given that these selection curricula have the ability to stratify trainees, early implementation may assist in career advise prior to applying for residency positions. Second, the study relied on proficiency cutoff scores derived from expert performance, rather than more junior trainees. The cut off scores from Ritter et al. were used because they are reported as time scores, allowing participants to easily determine their progress toward proficiency, which proved to be a source of motivation encouraging continued participation. The proficiency scores were not lowered, as we believed that the ability to acquire technical skills in a controlled environment devoid of surgical judgment is determined primarily by the opportunity to practice with adequate feedback, rather than stage of training. This approach is supported by the finding that most participants reached proficiency as determined by scores derived from expert performance levels, despite their status as pre-clinical medical students.

5.6 Conclusion

Trainees can be separated into performance groupings based on their ability to learn technical skills. Laparoscopic skills are more difficult to learn than open skills, with some trainees unable to reach proficiency in these, despite continued practice. Innate technical ability plays a substantially larger role in laparoscopy as compared to open skills, and this advantage persists
for the duration of the learning curve. Low performing novice trainees generally remain in this
tier across technical tasks. Furthermore, some of these lower tier performers never reach the
proficiency threshold, and demonstrate marked variability in their learning curves with no trend
towards a plateau phase even after forty repetitions in a training environment with structured,
standardized feedback. Given the increasing use of minimally invasive and endo-luminal
techniques in surgical practice, screening potential candidates for surgical training using
simulated tasks to identify the lowest tier performers may benefit both the students and training
programs. Although the findings of the present study take the recent literature further, a
longitudinal follow up of the lowest tier performers within the clinical environment would
further validate these findings.
Chapter 6: Randomized clinical trial to evaluate mental practice in enhancing advanced laparoscopic surgical performance

This chapter has been published as Louridas M, Bonrath EM, Sinclair DA, Dedy NJ, Grantcharov TP. Mental practice to enhance advanced laparoscopic surgical performance in the operating room: a randomized controlled trial. *Br J Surg.* 2015 Jan;102(1):37-44.

6.1 Abstract

**Background:** Mental practice, the cognitive rehearsal of a task without physical movement, is known to enhance performance in sports and music. Investigation of this technique in surgery has been limited to basic operations. The purpose of this study was to develop mental practice scripts, and to assess their effect on advanced laparoscopic skills and surgeon stress levels in a crisis scenario.

**Methods:** Twenty senior surgical trainees were randomized to either conventional training or mental practice groups, the latter being trained by an expert performance psychologist. Participants’ skills were assessed while performing a porcine laparoscopic jejunojejunostomy as part of a crisis scenario in a simulated operating room, using the Objective Structured Assessment of Technical Skill (OSATS) and bariatric OSATS (BOSATS) instruments. Objective and subjective stress parameters were measured, as well as non-technical skills using the Non-Technical Skills for Surgeons rating tool.

**Results:** An improvement in OSATS (P =0.003) and BOSATS (P =0.003) scores was seen in the mental practice group compared with the conventional training group. Seven of ten trainees improved their technical performance during the crisis scenario, whereas four of the ten conventionally trained participants deteriorated. Mental imagery ability improved significantly
following mental practice training (P =0.011), but not in the conventional group (P =0.083). No differences in objective or subjective stress levels or non-technical skills were evident.

**Conclusion:** Mental practice improves technical performance for advanced laparoscopic tasks in the simulated operating room, and allows trainees to maintain or improve their performance despite added stress.

### 6.2 Introduction

There has been a decrease in the total number of hours available for surgical training in many countries in recent years (Accreditation Council for Graduate Medical Education, 2013). These restrictions have been driven by societal concerns that fatigue may result in compromised patient care and decision-making, as well as efforts to limit work hours to acquire a healthier work–life balance. This has contributed increasing interest in training strategies outside the operating room to accelerate technical skills training and achievement of competent performance. Randomized trials have demonstrated that technical skills pre-training curricula result in trainees with a higher level of technical proficiency and a shortened learning curve in the operating room compared with those with no such training (Palter & Grantcharov, 2012; Scott et al., 2000; Scott & Dunnington, 2008; Zevin B et al., 2013; B. Zevin, Aggarwal, & Grantcharov, 2012). Technical skills warm-up before entering the operating room has also been shown objectively to improve skill during surgery (Calatayud et al., 2010; Chen et al., 2013; Kroft, Ordon, Arthur, & Pittini, 2012; J. Y. Lee et al., 2012). Both interventions require simulation technology and/or cadaveric specimens, with associated resource requirements, to run effectively. An inexpensive surgical enhancement strategy that improves operative skill, with minimal incremental resource requirements and without additional dedicated infrastructure, would therefore be useful. Mental practice, defined as the cognitive rehearsal of a task without physical movement, is used to improve performance in many fields. In sports psychology, positive effects of mental practice and mental preparedness on stress levels and performance during high-level competition have been demonstrated repeatedly (Burhans, Richman, & Bergey, 1988; Drisckell J., Cooper C., & Moran A., 1994; Paivio, 1985; Weinberg RS., Seaboume T., & Jackson A., 1992). Mental practice may confer benefits for the surgeon in terms of increased technical performance,
decreased stress levels, and improved decision-making skills in the operating room. Only a few studies, limited to basic laparoscopic operations performed by novice surgeons, have evaluated this approach as an adjunct to technical surgical skills training (S. Arora et al., 2010; S. Arora, Aggarwal, Sirimanna, et al., 2011). The aims of the present study were to develop a mental practice script for the performance of an advanced laparoscopic procedure, to assess the effectiveness of mental practice on advanced laparoscopic technical skill performance, and to determine whether it was associated with differences in stress levels and improvement in non-technical skills in a simulated crisis scenario.

6.3 Methods

The jejunojejunostomy (JJ) portion of a Roux-en-Y gastric bypass was selected as the advanced laparoscopic procedure for two reasons. Independent completion of the laparoscopic JJ portion of the Roux-en-Y bypass is recognized as a senior-level procedure for trainees, whereas assisting with the camera and retraction is considered a junior-level task. JJ can be realistically and reproducibly simulated in a box trainer using porcine bowel, and can be assessed using existing reliable and valid tools.

6.3.1 Mental practice script development

Interviews were conducted with subspecialty-trained bariatric surgeons to develop the mental practice script. Interviews continued until theme saturation of mental practice cues had been achieved. Interviews were recorded and surgeons asked to describe, in the first person, the visual cues (what you see) and kinesthetic cues (what you feel) at each step of a laparoscopic JJ. This technique was adopted from the performance psychology literature on script development and should not be mistaken for solely listing the operative steps (Callow & Hardy, 2004; Gregg MJ. & T., 2007; Holmes & Collins, 2001). Surgeons were asked to reflect back to each step of the operation to identify common pitfalls experienced by trainees. The voice recordings were transcribed, and iterative content analysis was used to identify emerging themes and create the
mental practice scripts. Pitfalls were added to the scripts if identified by more than one-half of
the expert surgeons.

6.3.2 Randomized trial

This study was a randomized single-blinded, two-armed trial conducted at a single large
academic institution in Canada. Before the start of the study the trial was registered through the
International Standard Randomized Controlled Trials Number (ISRCTN). Ethics approval was
granted by St Michael’s Hospital and the University of Toronto. Informed voluntary consent was
obtained from each study participant before randomization into either the mental practice or the
conventionally trained study arm using List Randomizer software (Randomness and Integrity
Services, Dublin, Ireland). No changes were made to the trial protocol during the course of the
study.

Postgraduate year 3 and 4 general surgery residents were eligible for recruitment to this study.
Participants who had completed fewer than five laparoscopic JJs in either the simulation
environment using a porcine bowel model in a laparoscopic box trainer, or in the operating room,
were excluded. Other exclusion criteria were: systemic illness affecting BP or heart rate (for
example, hypertension, diabetes mellitus or mood disorders), and use of prescription drugs that
modify cardiovascular response.

Previous work on the effects of mental practice on improving technical proficiency has shown an
average difference between trained and untrained groups of 5 points on the 35-point Objective
Structured Assessment of Technical Skill (OSATS) global rating scale (S. Arora, Aggarwal,
Sirimanna, et al., 2011). Analyses of these published data suggested a standard deviation for the
study sample of no more than 3.7. Using these findings, sample size calculation revealed the
need for at least nine individuals per group to detect a significant difference with an α of 0.05
and a power of 0.80.

All participants completed a demographic questionnaire at the start. Two objective validated
assessment tools were used to measure baseline mental imagery ability: the Mental Imagery
Questionnaire (MIQ) and the Movement Imagery Questionnaire Revised Second version (MIQ-
RS), which itself is broken down into visual imagery and kinesthetic imagery scores (S. Arora et al., 2010; Gregg, Hall, & Butler, 2010). To assess baseline technical skill, each trainee performed one laparoscopic JJ on a box trainer, using a porcine bowel model. The simulated laparoscopic JJ was video recorded for subsequent blinded rating using the OSATS (Martin et al., 1997) and the bariatric OSATS (BOSATS) scales (Zevin B et al., 2013). Finally, baseline stress levels were measured objectively by heart rate and BP (using a non-invasive automatic BP machine), and subjectively by the validated six-item State–Trait Anxiety Inventory (STAI) questionnaire for adults (Marteau & Bekker, 1992).

All subjects participated in a didactic lecture on the creation of a laparoscopic JJ. The lecture outlined the technical steps of the operation, with accompanying videos demonstrating the correct operative technique for each step. Following the lecture, a multiple-choice test was administered to ensure that all participants understood the steps of the operation. After the session, the instructional videos used in the didactic session were available to all participants for review at any time during the study. Residents in the intervention group additionally underwent mental practice training, consisting of in-person instruction from an expert performance psychologist and independent practice. Participants were taught first to perform a relaxation exercise with abdominal breathing, and then to begin mental practice guided by the mental practice scripts focusing on the kinaesthetic and visual cues required to ‘feel’ and ‘see’ each step of the laparoscopic JJ. Subsequently, all participants in the intervention group were provided with the written script, as well as a version of the technical skills videos from the didactic session that included a voice-over of the mental practice scripts. Each participant had 7 days to perform mental practice independently at home with the scripts and videos. To promote the use of mental practice, each trainee participated in three voice-recorded follow-up telephone calls in which they walked through the scripts verbally and received structured feedback from one investigator.

One week after baseline testing, both groups participated in a crisis scenario of a laparoscopic JJ procedure in the simulated operating room. The same porcine bowel model in a laparoscopic box trainer, as baseline, was used to allow blinding and direct comparative assessment before and after the intervention.

To create a realistic, standard operative environment, the roles of the anaesthetist, scrub nurse, circulating nurse and assistant surgeon were scripted and played by healthcare professionals who
were trained to react appropriately within the script guidelines. As the trainee performed the procedure, at a set operative step, the simulated patient unexpectedly had an anaphylactic reaction to the routinely administered preoperative antibiotics. The crisis was introduced purposefully when the gastrointestinal tract was interrupted, driving the resident to manage the crisis then finish the operation rather than abort the procedure. Immediately following participation in the crisis scenario, participants were asked individually to keep all information concerning the session confidential, and all agreed. Three different crisis scenarios were prepared in case of a breach of confidentiality. Only a single scenario was used, however, as participants continued to be surprised and stressed as measured by objective and subjective parameters, with no indication of any breach of confidentiality.

To ascertain objective performance, the video recordings of the laparoscopic JJs were assessed using both OSATS and BOSATS scales. All laparoscopic-view videos (baseline and simulation crisis JJ) were edited to start when the laparoscopic graspers entered the box trainer and to end after closing the common enterotomy. The pause reflecting the crisis period was removed from the laparoscopic video before review in random order to ensure that the raters were blinded to both the participant and study stage (baseline versus after intervention).

Stress levels were assessed using both objective and subjective instruments. During the simulation, BP was taken at five fixed points in the operation using an automatic BP cuff, and heart rate was recorded at 1-s intervals, using a non-invasive chest strap monitor (Polar Electro, Kempele, Finland). Participants wore monitors for the whole of the scenario, allowing parameter measurement with minimal interruption to the surgical task being performed. Subjective stress was scored by each participant using a validated six-item version of the Spielberger STAI (Marteau & Bekker, 1992), which was administered both before and after the simulation.

The full simulated surgical scenario was video recorded for each participant to assess non-technical skills. Two trained raters used the Non-Technical Skills for Surgeons (NOTSS) scoring system to evaluate the trainees (Yule et al., 2008). Raters were non-surgeons who had received dedicated training from a surgeon researcher experienced in the assessment of surgeons’ non-technical skills within the operating room. This training included instruction and training on roles, responsibilities and behaviours within the operating room, multiple case-based training sessions, practice ratings using intraoperative videos and rater calibration by comparing scores.
To ensure objective assessment, raters were blinded to the purpose of the study and group randomization of the participants, had no association with the general surgery training programme, no personal or professional relationship with the study participants, and were not associated with the development or implementation of this research.

To assess participant satisfaction with the use of mental practice, all intervention group residents completed a short questionnaire to gauge whether it was considered a valuable adjunct to surgical training.

6.3.3 Statistical analysis

Descriptive statistics were calculated for all variables. Non-parametric statistical tests were used to compare technical, non-technical skills and STAI scores within and between groups (Wilcoxon rank sum test and Mann–Whitney U test). To account for the difference in baseline scores, change scores (baseline score – final score) were used to compare the cohorts. Parametric tests (independent and paired t tests) were used to evaluate differences in heart rate and BP.

6.4 Results

6.4.1 Mental practice scripts

Saturation of themes was reached after eight bariatric surgeons had been interviewed. Six operative substeps were identified for the laparoscopic JJ, with each broken down into visual and tactile cues. For example, a portion of the script of step 1 ‘running the proximal limb of small bowel 50 cm from the ligament of Treitz’, read:

Visual: ‘I grasp the bowel whilst I see it; I never grab the bowel with the graspers out of sight. I don’t want to see any blotchy haematomas, or excessive blanching where I’ve grasped. These are visual signs of trauma: which means, I’ve grasped too hard or incorrectly’.
Kinaesthetic: ‘I try to feel where the bowel wants to go. I don’t push the bowel or direct it
to an area where I can feel tension. I run the bowel with zero tension in a clockwise
fashion. If the bowel doesn’t want to come, something’s wrong. It’s probably stuck
somewhere or I haven’t got a good grab’.

A minimum of two to a maximum of five pitfalls was identified for each step of the operation.
An example of a pitfall repeatedly identified by experts during step 1 was ‘working off screen’:

Working off screen: ‘Trainees may grab the bowel when it’s out of sight and therefore
they cannot see if they are applying too much force or have caused injury. This can result
in unrecognized bowel injury’.
Figure 14: CONSORT diagram illustrating progress through the phases of the study

6.4.2 Randomized trial

The progress of participants through the phases of the study is shown in Figure 14. Demographic characteristics of the groups were similar, but laparoscopic experience was not the same because the intervention group had performed more laparoscopic procedures as the primary surgeon and
more laparoscopic JJs in the operating room and simulation centre (Table 19). The OSATS and BOSATS scores at baseline were higher in the intervention group than those in the control group. Change scores (final score – baseline score) were therefore used to compare the cohorts. At baseline, all remaining measurements were equal between groups including mental imagery ability, and objective and subjective stress parameters (Table 20).

Table 19: Demographics of study participants

<table>
<thead>
<tr>
<th></th>
<th>Conventional training ($n = 10$)</th>
<th>Mental practice ($n = 10$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postgraduate year</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Sex ratio (M : F)</td>
<td>9 : 1</td>
<td>7 : 3</td>
</tr>
<tr>
<td>Handedness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Left</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Completed FLS training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>No</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>No. of junior-level laparoscopic procedures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 10</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>≥ 10</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>No. of procedures as primary surgeon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 10</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>≥ 10</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>No. of laparoscopic JJs in OR and simulation centre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 10</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>≥ 10</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

FLS, Fundamentals of Laparoscopic Surgery; JJ, jejunojenustomy; OR, operating room.

Table 20: Result of baseline assessments of technical skill and mental rotation ability

<table>
<thead>
<tr>
<th></th>
<th>Conventional training</th>
<th>Mental practice</th>
<th>$P^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical skill score (points)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSATS</td>
<td>18 (17, 21)</td>
<td>22 (21, 27)</td>
<td>0.043</td>
</tr>
<tr>
<td>BOSATS</td>
<td>21 (20, 32)</td>
<td>28 (26, 35)</td>
<td>0.123</td>
</tr>
<tr>
<td>Mental rotation ability score (points)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIQ</td>
<td>32 (27, 39)</td>
<td>33 (26, 39)</td>
<td>0.912</td>
</tr>
</tbody>
</table>
RVMIQ  43 (41, 45)  46 (35, 49)  0.684  
RKMIQ  40 (36, 44)  42 (33, 48)  0.971

Values are median (i.q.r.). OSATS, Objective Structured Assessment of Technical Skill; BOSATS, bariatric Objective Structured Assessment of Technical Skill; MIQ, Mental Imagery Questionnaire; RVMIQ, Revised Vividness Mental Imagery Questionnaire; RKMIQ: Revised Kinaesthetic Mental Imagery Questionnaire. *Mann–Whitney U test.

Greater improvements in technical skills were seen among patients who underwent mental practice training, with significantly higher OSATS and BOSATS change scores than among those who received conventional training. Seven of ten participants in the intervention group improved their OSATS score by at least 5 points on the final crisis scenario, and the remaining three improved by 0–4 points. None of these participants deteriorated in skill when placed in the crisis environment. In contrast, only two of the ten conventionally trained residents improved their OSATS score by 5 points or more, four improved by 0–4 points and four deteriorated in skill when tested in the crisis environment (Figure 15 and Figure 16). Those in the intervention group had significantly greater median absolute OSATS and BOSATS scores on final assessment than at baseline. These changes were not seen in the conventional group (Table 21).
Figure 15: Comparison of Objective Structured Assessment of Technical Skill (OSATS) change scores between groups. Median (line within box), interquartile range (box), and range (error bars) excluding outliers (circles) are shown. Dotted line indicates baseline performance. $P = 0.003$ (Mann–Whitney U test)
Figure 16: Comparison of bariatric Objective Structured Assessment of Technical Skill (BOSATS) change scores between groups. Median (line within box), interquartile range (box), and range (error bars) excluding outliers (circles) are shown. Dotted line indicates baseline performance. $P = 0.003$ (Mann–Whitney U test)
Table 21: Technical skill results at baseline and following training

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Crisis scenario</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conventional training</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSATS score (points)</td>
<td>18 (17, 21)</td>
<td>19 (16, 25)</td>
<td>0.734</td>
</tr>
<tr>
<td>BOSATS score (points)</td>
<td>21 (20, 32)</td>
<td>24 (20, 30)</td>
<td>0.622</td>
</tr>
<tr>
<td><strong>Mental practice</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSATS score (points)</td>
<td>22 (21, 27)</td>
<td>30 (25, 36)</td>
<td>0.005</td>
</tr>
<tr>
<td>BOSATS score (points)</td>
<td>28 (26, 35)</td>
<td>37 (35, 40)</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Values are median (i.q.r.). OSATS, Objective Structured Assessment of Technical Skill; BOSATS, bariatric Objective Structured Assessment of Technical Skill. *Wilcoxon signed-rank test.

Residents in the mental practice group demonstrated a significant improvement in median MIQ scores following training. No significant difference in median scores was identified in the conventional training group. No significant differences in mental imagery abilities were identified for either study arm, using the MIQ-RS (Table 22). Systolic BP, diastolic BP and heart rate increased significantly during the crisis scenario in all participants compared with baseline: mean (s.d.) increase 18(25) mmHg ($P = 0.004$), 20(13) mmHg ($P < 0.001$) and 14(14) b.p.m. ($P < 0.001$) respectively. During the crisis moment the intervention and control groups experienced equally high systolic BP (139(19) and 137(17) mmHg; $P = 0.830$) and diastolic BP (94(17) versus 94(10) mmHg; $P = 1.000$). The groups reported feeling equally stressed during the crisis scenario as measured using the STAI score: median 11 (i.q.r. 10–14) versus 12 (11–13) respectively ($P = 0.853$).

There was no significant difference in non-technical performance when mean values of the four categories of NOTSS were compared between groups ($P = 0.853$). Inter-rater agreement was excellent, with an intraclass correlation coefficient of 0.80.
Eight of ten trainees in the intervention group felt that mental practice should be incorporated in the training curriculum and seven of ten believed they would use this technique to prepare for operations in the future.

Table 22: Mental imagery ability at baseline and following training

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Crisis scenario</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIQ score (points)</td>
<td>32 (27, 39)</td>
<td>42 (27, 45)</td>
<td>0.083</td>
</tr>
<tr>
<td>RVMIQ score (points)</td>
<td>43 (41, 45)</td>
<td>43 (39, 49)</td>
<td>0.270</td>
</tr>
<tr>
<td>RKMIQ score (points)</td>
<td>40 (36, 44)</td>
<td>39 (36, 47)</td>
<td>0.670</td>
</tr>
<tr>
<td>Mental practice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIQ score (points)</td>
<td>33 (26, 39)</td>
<td>41 (31, 43)</td>
<td>0.011</td>
</tr>
<tr>
<td>RVMIQ score (points)</td>
<td>46 (35, 49)</td>
<td>43 (40, 47)</td>
<td>0.757</td>
</tr>
<tr>
<td>RKMIQ score (points)</td>
<td>42 (33, 48)</td>
<td>44 (35, 47)</td>
<td>0.234</td>
</tr>
</tbody>
</table>

Values are median (i.q.r.). MIQ, Mental Imagery Questionnaire; RVMIQ, Revised Vividness Mental Imagery Questionnaire; RKMIQ, Revised Kinaesthetic Mental Imagery Questionnaire. **Wilcoxon signed-rank test

6.5 Discussion

Mental practice with scripts and accompanying voice-over instructional videos significantly improved mental imagery ability as well as advanced laparoscopic technical skill.

Somewhat unexpectedly, mental practice had no effect on measured stress level or non-technical skills.

Mental practice not only improved advanced laparoscopic performance but technical performance also improved, despite the added stresses of the unexpected intraoperative crisis. Decreased performance following stressful situations secondary to an involuntary physiological surge of cortisol, with an increase in heart rate and BP, is well recognized (Sonal Arora et al., 2009; Kirschbaum, Wolf, May, Wippich, & Hellhammer, 1996). None of the trainees who
received mental practice exhibited deteriorating performance scores from baseline during the crisis, whereas four of the ten conventionally trained had worse performance. This finding is important for surgical trainees who work routinely under pressure, and undoubtedly will be required to perform under stress during their surgical careers. Previous studies (S. Arora, Aggarwal, Sirimanna, et al., 2011; Geoffrion et al., 2012) have tested participants directly after in-person mental practice coaching. The present study demonstrated the effectiveness of this technique with delayed testing after mental practice at home, which may be a more realistic approach to implementation of mental practice into residency training.

The effectiveness of mental practice in improving performance is dependent on achieving improvements in mental imagery ability (Sevdalis, Moran, & Arora, 2013). Although it is difficult to quantify and measure participants’ use of mental practice techniques directly, tools to assess mental imagery ability, such as the MIQ, may be able to identify improvements associated with their use. In the present study, participants demonstrated a significant improvement in MIQ scores following mental practice training but not after conventional training. Surprisingly, stress levels and non-technical skills were similar in the two groups. This was in contrast to one other study (S. Arora, Aggarwal, Moran, et al., 2011) in which mental practice significantly decreased stress scores measured by heart rate, STAI and urinary cortisol in novice surgeons performing a simulated laparoscopic cholecystectomy. Reasons for similar responses between groups in the present study may be related to the mental practice scripts being highly focused on kinaesthetic and visual imagery for technical skill, rather than motivational enhancement imagery that focuses on optimizing performance during times of increased pressure or stress. Although the intervention group was able to outperform the conventional trainees in the presence of significant stress, similarities between the groups with respect to non-technical skills suggest that altering behaviours such as communication and team management require dedicated non-technical skills training30, rather than an indirect intervention such as mental practice or progression of increased experience during surgical training (Alvand et al., 2012; Flin et al., 2007).

This study has a number of limitations. Baseline OSATS scores were significantly different between the groups. This difference was probably explained by the study participants’ variable exposure to minimally invasive surgery rotations before the start of the study. The use of change scores in the analysis should, however, have overcome most of this effect. Non-technical skills raters were not surgeons, which may have impeded their ability to discern differences in non-
technical skills associated with crisis management. These individuals had, however, received extensive instruction and training in non-technical skills assessment before the start of the study, minimizing the impact of their lack of direct surgical experience. It is possible that there might have been differences in the relative contribution of the different components of the mental practice intervention (in-person training, independent practice with written script, review of performance videos with audio or written script) to the observed differences in technical performance, but the study design did not allow assessment of these potential differences. Cortisol levels were not used to quantify stress because the majority of trainees had clinical duties to attend to, and it was not feasible for them to participate in the study while avoiding the many confounders of cortisol levels (Kelly, Young, Sweeting, Fischer, & West, 2008).

Mental practice as an adjunct for training of advanced laparoscopic skills appears to be an effective method of improving performance in surgical trainees, with the added stressor of a crisis scenario in the simulated operating environment. These findings suggest that it may be a promising adjunct to surgical training programmes.
Chapter 7: General discussion

7.1 Thesis summary

Work hour restrictions, stringent patient safety precautions, and the incorporation of minimally invasive surgical techniques into training have made it increasingly difficult for residents to reach technical competence (Levine & Spang, 2014; Watson, Mathew, & Williams, 1995). Therefore, optimizing the selection of surgical residents and incorporating performance adjuncts into training may help ensure technical competence is consistently met. Thus the focus of this thesis was to investigate different methods of screening for technical ability at the time of selection of incoming surgical trainees by first completing a systematic review on surrogate markers that may predict technical skill. This review demonstrated that visual spatial tests may have the ability to predict laparoscopic performance in the novice trainee. To this end, I then completed an original study to assess whether visual spatial tests predict laparoscopic performance of incoming surgical trainees, however no reliable association was found. Therefore, my second original study was a national Delphi questionnaire of Canadian GS PDs to determine which simulated technical skills would be most appropriate to screen incoming residents. Both open and laparoscopic skills met expert consensus. My third original study incorporated these recommendations and assessed medical students’ learning curves of open and laparoscopic tasks over a one month training curriculum starting with basic then advanced technical skills on low fidelity simulation models. Students acquired technical skills at variable rates and a small subset (8-15%) were low performers who were unable to reach proficiency in any of the laparoscopic tasks. Lastly, to explore in-training adjuncts for technical performance, a randomized controlled trial was completed to assess whether a cognitive MP could be used to improve technical performance of senior surgical residents during advanced laparoscopy, during a simulated crisis scenario. The MP group outperformed the control group for technical skill but no difference was demonstrated between the groups for non-technical performance or measured stress levels.
Overall the work in this thesis, including one systematic review and four original manuscripts, demonstrated that a simulation surgical skills curriculum assessing progression of skill acquisition may be an effective method for screening technical ability of incoming trainees. Furthermore, MP is an affective adjunct to improve technical performance in senior residents performing advanced laparoscopy and therefore should be considered for incorporation into surgical training programs.

7.2 Findings with respect to the aims of the thesis

At the time of selection

7.2.1 To evaluate whether previous surgical experiences, non-surgical experiences and 2D-3D visual spatial tests correlate with laparoscopic skills in the novice surgical trainee

In the context of finding a screening test for incoming surgical trainees that has the ability to predict technical performance, surrogate markers and background characteristics were investigated first because both these components could easily be added to the existing selection process. As outlined in section 2.1, the selection process already requires that applicants submit a standardized curriculum vitae, followed by participation in an in-person interview. Therefore, if background characteristics were able to predict technical skill, they could be added to the existing written application. In addition, the visual spatial tests studied were either paper-based or computer-based and could also be added to the existing in-person interview process with minimal additional resources.

A systematic review was performed to assess the existing evidence supporting the use of surrogate makers and their ability to predict technical skill (section 2.3.2). Video game experience, time spent in the real operating room and visual spatial tests were the only markers found to have a higher number of studies demonstrating a positive association with technical performance (Dimitriou et al., 2009; Hislop et al., 2006; Kolozsvari et al., 2011; Madan et al., 2005; Nomura et al., 2008; Paschohl et al., 2011; D. Stefanidis, Korndorfffer Jr, et al., 2006; Tangchitnob et al., 2011; Van Hove et al., 2008). However, video game experience is a problematic screening question because the reported gaming advantage is likely due to the learnt
process of responding to visual cues from a 2D screen which can then be transferred to laparoscopy (Dimitriou et al., 2009; Paschold et al., 2011). Thus video game experience likely does not evaluate technical aptitude but rather a learnt skill. Similarly, time spent in the real operating room is not a fair selection question, because it does not address technical aptitude but again, learnt skill. Furthermore, time spent in the operating room will differ between applicants due to their medical school program structure, the number of weeks of surgery they are exposed to and the number of learners in the operating room. Although these external factors directly influence this screening question it will have no bearing on true technical ability. Therefore, these two background factors are not suitable for screening incoming trainees.

However, finding a background characteristic or experience that may be associated with technical performance would be advantageous to the existing selection process due to ease of implementation. Therefore, background experiences were explored in both original studies of incoming surgical trainees and medical students in this thesis (chapter 3 and 5). The background experiences were expanded to include involvement in sports, music, gaming and characteristics such as gender, handedness etc. but unfortunately none of these demonstrated a consistent positive association with technical performance. Despite my persistent effort to search for a background characteristic or experience to incorporate into selection to screen for technical aptitude, none reliably demonstrated a positive association with technical ability. Instead, background experiences may be helpful to determine other desirable attributes for incoming trainees (e.g. ability to work in a team, work ethic, interest in research) and further study in this area may prove fruitful.

Twenty-five visual spatial tests were identified in the literature search but only three were thought to have some evidence of predicating laparoscopic performance: PicSOR, cube comparison tests and card rotation test. Of the three, PicSOR demonstrated the most consistently positive association with simulated laparoscopic skills in a box trainer and virtual reality simulator (Buckley et al., 2013; Buckley et al., 2014; A. G. Gallagher et al., 2003; Kolozsvari et al., 2011; McClusky et al., 2005). However, often these associations were of subcomponents of the scoring metrics, rather than overall performance. Scores on the cube comparison and card rotation tests had equal numbers of studies that demonstrated positive and negative associations with laparoscopic performance, therefore it was still unknown whether these tests would be useful for the purpose of selection (Buckley et al., 2013; Buckley et al., 2014; L. Enochsson et al.,
2006; Groenier et al., 2014; Kolozsvari et al., 2011; McClusky et al., 2005; Nugent et al., 2012; Dimitrios Stefanidis et al., 2007). To further understand this relationship, these three visual spatial tests were selected and compared to laparoscopic camera navigation on the VR simulator and laparoscopic circle cut on the box trainer.

Of the three visual spatial tests listed above, only the cube comparison test demonstrated a positive association with overall technical performance on the VR LCN task. No other positive associations were noted. Although disappointing that these tests were not able to screen for technical ability of novice trainees, it speaks to the complexity of predicting technical aptitude. Visual spatial tests may play a role in laparoscopy, however in isolation it does not explain the reason for variable performance between trainees. Technical performance is likely a multifaceted process involving many interacting abilities, which to date are not fully understood.

7.2.2 To solicit program directors’ opinions on the proportion of trainees who do not achieve the minimum technical standards expected at the time of graduation

To assess whether there is a potential for improving the selection process by screening for future technical ability, current PDs of Canadian GS training programs were surveyed. Specifically, they were asked whether they felt some contemporary GS trainees are unable to reach technical competence despite completing a full post-graduate GS training program. The majority of PDs reported that some trainees do not achieve the minimum technical standard at the time of graduation, with the stated proportion ranging from 5-15%. Although no historical Canadian data are available for comparison, this number is consistent with previously published findings from a number of European centers which reported that 5-20% of their participants had difficulty reaching technical competence (A. Alvand, S. Auplish, T. Khan, et al., 2011; Cushchieri, 2003; Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004).

In 2015, 87 trainees entered GS training across Canada. Over the past 5 years, the number of available national training positions in GS has ranged from 86-95. Consequently, if the 2015 numbers are extrapolated to reflect 10-15% of trainees, nationally 4-14 incoming residents per year are at risk of not reaching technical competence. These numbers represent a proportion of trainees that have successfully gone through the full selection process. They have worked with a
number of GS staff, obtained reference letters from surgeons in the field, received multiple
evaluations and have interviewed to be matched into competitive positions. This information
further supports that technical aptitude screening is an area within the selection process that has
room for improvement.

7.2.3 To establish a national consensus on the desired attributes of GS
candidates, and the technical skills that would be most indicative of
future performance

Clear consensus was met amongst the PDs on a number of non-technical attributes that are
desirable for a GS trainee and were prioritized above technical skill. The attributes that received
the highest level of agreement included: work ethic, passion for GS, professionalism, ability to
work in a team, sound judgment, ability to make decisions and the ability to assimilate
information to formulate an opinion. Therefore, it was clear from the PDs that although technical
ability is required for surgical training its importance was second to the non-technical skills listed
above. GS is a team based interdisciplinary surgical field where nurses, allied health members
(e.g. occupational therapist, physiotherapist, wound care specialists and dieticians), patients and
their families are imperative to each patient’s peri-operative care. Therefore, it is understandable
that non-technical skills were emphasized by the PDs as most important. Selecting residents who
will excel in these qualities is of utmost importance and future research is needed to identify
screening tools that predict these attributes.

However, even though technical skill was not considered the most important attribute, it also
reached positive expert consensus. Furthermore, consensus was reached on two basic open tasks
(one-handed tie and subcuticular suture) and two laparoscopic tasks (coordination, grasping and
cutting). The tasks that reached consensus were all basic skills that are required across all GS
disciplines, whether applying to a rural, community or urban training environment or future
surgical practice. Furthermore, the simulation equipment required to implement these tasks is
readily available in all basic simulation centres with no specialized equipment required.

The reason expert consensus was used to select technical tasks is due to the paucity of
information available in the literature to select evidence based technical tasks for selection.
Furthermore, many questions remain unanswered, including what the appropriate standards are
for each task, how well these tests discriminate between technical performance and at what time point(s) in the learning of these skills reflects future clinical performance. Therefore, although expert consensus was used as a starting point to choose technical tasks, original research is needed to delineate these details.

7.2.4 To quantify different learning patterns among trainees for both basic and more advanced laparoscopic and open skills

The existing surgical literature had examined the LCs of single simulated laparoscopic tasks or procedures, (Grantcharov & Funch-Jensen, 2009; M. P. Schijven & Jakimowicz, 2004) but had not determined the learning patterns of tasks of varying difficulty. Furthermore LC data for open tasks to discriminate different learners had not been reported in the literature.

In today’s surgical training environment, at the time of graduation residents are expected to be component in both open and laparoscopic procedures. With that said, subspecialty training within GS allows each individual to tailor their surgical practice to encompass technical elements that complement their strengths and interests (e.g. breast surgery requires predominately open skills vs bariatric and foregut surgery which require advanced laparoscopic skills). Thus, understanding learning curve patterns across different tasks within open and laparoscopic skill may be helpful in understanding where low performers are most disadvantaged.

Overall, learning patterns ranged from very efficient learning, demonstrated by a steady progression towards competence, to, in contrast, widely variable learning, where improvement was incrementally small due to highly fluctuant performance. Furthermore, as the task increased in difficulty, variability in performance increased as well. Trainees took longer to reach competence and found it to be more difficult to maintain consistent performance at this level. Therefore, trainees and faculty should be aware that fluctuating performance is expected as task difficulty increases and increased practice will be required to maintain competence performance. Simulation training is likely most beneficial for the fluctuating portion of the learning curve of these tasks.

More specifically, four distinct learning clusters were identified within these general learning patterns. Clusters 1-4 grouped individuals over the range from high to low performers,
respectively. Low performers had persistent difficulty with laparoscopic tasks as compared to open tasks, and also demonstrated the most significant amount of fluctuation between repetitions.

7.2.5 To determine whether trainees stay within their learning patterns across simulated tasks of varying difficulty (basic and advanced) and type (minimally invasive and open)

The study described in chapter 5 demonstrated that there was a significant association between individuals to stay within their assigned cluster with lateral movement to a neighboring cluster. Thus, despite the tasks increasing in difficulty, low performers demonstrated poor acquisition of technical skill within laparoscopy as compared to their peers, even when learning the easier tasks.

This finding is interesting because anecdotally it may be hypothesized that all trainees will be able to learn easy tasks but not all trainees will be able to learn more difficult tasks. However, this was not the case. Instead, trainees who demonstrated difficulty with basic tasks had even more difficulty with advanced tasks. Furthermore, this provides evidence that screening with basic laparoscopic skills may a sufficient discriminatory test to differentiate technical aptitude for incoming trainees.

7.2.6 To identify a subset of trainees who consistently fail to reach proficiency on simulated tasks and determine the features of their learning curves that separate them from their peers

The distinct learning features of low performing trainees were unique to laparoscopic tasks and included high variability between repetitions, inability to reach the proficient cut off score and no true plateau phase in the learning curve. These learning curve features were demonstrated in the simulation center, which is a very controlled environment when compared to the operating room. The equipment, instrument set-up and simulation models were constant throughout the learning period and despite this, variable performance was demonstrated from the lowest performers. The real operating environment is far more complex. Patient factors include tissue integrity, set-up, body habitus and anatomy different from case to case. Therefore, it may be reasonable to infer that these factors would further exacerbate already variable technical performance.
Furthermore, the question whether continued practice would eliminate these concerning learning curve features is unresolved. So far in the surgical education literature, 10-30 repetitions have been used to quantify learning curves of simulated tasks. The study in chapter 5 increased the practice window to 40 repetitions, and when comparing the results of each task a 5-20% low performing group was consistently present. Furthermore, current surgical trainees are expected to reach technical competence within an environment with work hour restrictions, increasing patient safety precautions and where surgical techniques are becoming more difficult. Acquiring technical skill at a reasonable rate is necessary to progress through training in a timely manner. In addition, technical performance is only one of the many required competencies set by the Royal College of Physicians and Surgeons. Therefore, dedicated time is required to also grow the non-technical elements of becoming a safe surgeon. Although the accepted time to acquire a technical skill to competence has not been formally defined within the literature, forty repetitions is the longest training window that has presently been examined in the literature. The associated learning curve features demonstrated by low performing trainees provide some evidence that proficiency will not be met in a timely manner, as progress toward the proficiency threshold was not convincing.

**During training**

7.2.7 To develop a MP script for the performance of an advanced laparoscopic procedure

A mental practice script was created primarily focusing on CS mental practice as compared to MG or MGA. CS mental practice was chosen because it is recommended when trainees are beginning to learning new skills, pertinent to the study population of interest, in comparison to expert surgeons who would likely benefit from a performance enhancement script that focuses on MG or MGA.
7.2.8 To assess the effectiveness of MP on advanced laparoscopic technical skill performance

The effectiveness of mental practice was explored as a technique to be incorporated into surgical residency specifically as an adjunct to improved technical performance in senior level trainees. In the RCT described in chapter 6, the trainees who participated in MP technically outperformed the control group. As described in section 2.6.2, MP focuses on the cognitive process of learning an operation as compared to physical practice. To date, simulation curricula have largely focused on physical practice, i.e. performing a task and learning through repetitive practice. However, although simulation has advanced tremendously over the years, creating a realistic simulation of full laparoscopic operations or even a portion of an operation, is challenging. Trainees can practice operations on a VR simulator, however senior level residents who have experience operating with real tissue and real surgical equipment do not respond well to VR simulation due to the limitations of the technology to create a realistic operative feel (Shetty, Zevin, Grantcharov, Roberts, & Duffy, 2014). Furthermore, to create a realistic operation in a box trainer requires a high fidelity model (i.e. cadaveric porcine bowel), which is expensive and often limited to one time use. Therefore, instead of relying on physical practice alone in senior trainees, MP can be used as an adjunct to enhance technical performance in the real operating room.

7.2.9 To determine whether MP is associated with differences in stress levels and improvement in non-technical skills in a simulated crisis scenario.

Mental practice training was not associated with differences in stress levels or improvement in non-technical skills in a simulated crisis scenario. The reason MP was not associated with decreased stress level may be due to the way stress was measured. The non-invasive physiologic stress levels used included blood pressure and heart rate, with a higher heart rate and blood pressure indicative of increased stress. However, all surgical trainees entering the simulated operating room scenario had heart rate and blood pressure levels above their baseline and the incremental difference did not differ between the groups. More sensitive and specific measures of stress such as heart rate variability or salivary cortisol may have detected a difference, however were not available (Crewther et al., 2016; Kirschbaum et al., 1996). In addition, the
scripts were focused on the technical component of the operation, which is known as CS mental practice as compared to MG or MGA. MG and MGA MP have a stronger focus on relaxation and coping during stressful moments. Therefore, training the participants with these MP techniques may have shown a difference in stress level and perhaps also in non-technical skill.

7.3 Limitations

Each chapter’s full text manuscript includes a limitations section specifically pertaining to the study. However, in this overall limitations section I will focus on three broader concepts that influence surgical education research in the area of technical skill assessment in both selection and in-training: 1. defining cut off scores for competence, 2. assessment of simulated and real operative technical skills using GRS and 3. adequate sampling.

7.3.1 Defining cut off scores

The methodology portion of this thesis that investigated participants’ learning curves required the use of a defined threshold for competence in technical performance. However, this work is limited by the fact that it remains unclear within the surgical education literature what the optimal cut off scores should be for different technical task performance levels and assessment tools. Some reason for this lack of clarity may be the simplified methodology used by studies to set these cut off scores, the inconsistent terminology used within the literature to describe different thresholds of performance, and the inability to transfer cut off scores across studies.

A common methodology for setting cut off scores described in technical skills simulation studies is the use of expert surgeons’ performance as a benchmark. For example, this approach consists of having an expert surgeon complete five to ten repetitions of a given task, and defining the cut off score as one or two interquartile ranges or standard deviations worse than the median or mean of their performance across all repetitions (E. M. Ritter & Scott, 2007; Zevin B et al., 2013). There are many problems with this approach. First, the cut off score depends on the performance of a limited number of experts who may not represent the population norm for the accepted
safety threshold. Second, the selected experts are often subspecialty trained surgeons (i.e. experienced laparoscopic surgeons for laparoscopic skill or experienced open surgeons for open skills). Therefore, their performance scores may be consistently far above the competency threshold, closer to a proficient or expert level of performance.

Within the surgical education literature, the words competent and proficient are often used interchangeably, despite these terms describing two distinct performance levels. When these terms are linked back to their origin in education theory, they are meant to describe different phases of skills acquisition. Dreyfus and Dreyfus (1980) outline a five-stage model of the mental activities involved in directed skill acquisition (Dreyfus & Dreyfus, 1980). A student will begin as a ‘novice,’ and at this stage of learning the trainee is dependent on the rules and only feels responsibility to follow the rules. With practice and experience the student becomes an ‘advanced beginner’ where he/she identifies conditional rules but continues to only feel a sense of responsibility within the rules. The next phase is termed ‘competent,’ and is when the student feels responsible for making decisions and begins to sort information by importance. The forth state is ‘proficient’ whereby the student feels even more responsibility for his/her actions and is able to use pattern recognition to assess what to do. Lastly ‘expert’ level is reached as the skills become intuitive or automatic (Figure 17). Therefore, using the terms competent and proficient interchangeably is incorrect and confuses different stages of learning. Discriminating these levels may be helpful in setting cut off scores for trainees of different levels and may shield against setting scores that are too high or too low.
Figure 17: A five-stage model of the mental activities involved in directed skill acquisition (adapted from (Dreyfus & Dreyfus, 1980)).

Therefore, using methodology that is able to discriminate between competent and proficient may help with this long-standing limitation of setting cut off scores.

Recently, standard setting methodology has been suggested as an alternative approach to setting cut off scores. Within medical education, standard setting has been used to set pass/fail scores for high-stakes assessments for written, oral and OSCE examinations (McKinley & Norcini, 2014; Norcini, 2003), however this methodology is new to technical skills assessments.

Three standard setting methodologies have been demonstrated to appropriately set pass/fail scores for technical skills of surgical trainees. These include: contrasting groups, borderline group and borderline regression methods (Figure 18 and Figure 19) (de Montbrun, Statterthwaite, & Grantcharov, 2015). Contrasting groups is centered on the idea that within a given population there are trainees that will undoubtedly pass or fail a task (Livingston & Ziesky, 1982). For example when performing a technical task an expert examiner will score the trainee’s overall performance as competent or not competent. Plotting histograms of the scores (either GRS or checklist scores) in these two categories will result in an intersection between the groups
which is determined as the passing score (Norcini, 2003)(Figure 18 a). Borderline group methodology is centered on the idea that the pass/fail score should be set at the level of the borderline student (Sturmberg & Hinchy, 2010). The borderline student is defined as the individuals sitting on the edge of the passing score and this score becomes the pass/fail score (Figure 18b). For example when the student is performing a technical task the expert examiner will score their performance as neither competent or not competent but borderline.

Borderline regression is also centered on the idea of the borderline candidate (Sturmberg & Hinchy, 2010). However, instead of using a subset of scores for the calculation of the pass fail score, a linear regression analysis is used (Figure 19).

![Diagram of Standard Setting Using a. Contrasting Groups and b. Borderline Group Methodology](images/adapted_from_de_montbrun,_statterthwaite,___grantcharov,_2015)

Figure 18: Standard setting using: a. contrasting groups and b. borderline group methodology (images adapted from (de Montbrun, Statterthwaite, & Grantcharov, 2015))
Figure 19: Standard setting using borderline regression methodology (image adapted from (de Montbrun et al., 2015))

However, using standard setting methods to set cut off scores does not imply that these scores will be transferable to other studies. Fraser et al. set passing scores using the contrasting groups methodology (Figure 18a). The contrasting groups were defined as senior residents and junior trainees and the interception of their scores was termed the competence threshold (Fraser et al., 2003). However, despite using one of the described standard setting methodologies, the reported cut off ‘competency’ scores in this study seem to be very low compared to the proficiency threshold set by Ritter et al. Reasons for these low scores could be that the residents who participated in the study may have had minimal experience with the FLS tasks resulting in low performance scores, which then shifted the competency threshold downward, affecting their transferability to other studies.

When seeking cut off scores for chapter 5 of the present work, the cut off scores reported by Fraser et al. were found to be too low for the participants. The medical students in cluster 1 and 2 (top and high performers) surpassed the cut off scores set by Fraser et al. within their first 2-5 repetitions of the FLS tasks, yet it was obvious to the instructors that they were not comfortable performing the task and certainly far from competent in performing them.

Therefore, there is a need within the field of surgical education to develop and define cut off scores that are reliable and credible, to use consistent terminology when referring to the
7.3.2 The implementation of global rating scales as a routine assessment method during selection and surgical training

Global rating scales are considered the gold standard in assessing technical performance during procedures or operations, and are believed to be superior to checklists due to their ability to discriminate between performance quality (Regehr et al., 1998). However, the inclusion of these assessment tools in surgical curricula and ongoing surgical training is limited due to substantial time and resource requirements associated with their regular and objective use.

Watching each operation to determine a GRS performance score is extremely time consuming. To use the GRS assessment tool correctly, the assessor is required to watch the procedure in normal playback speed or in real time. This method was used in chapter 5 for the two open tasks, which were graded using the mOSATS GRS. The present work required the assessment of 5200 videos, which resulted in 260 hours of rating time. In the context of a single study, this approach was feasible. However, if implemented as a routine assessment tool within an ongoing educational curriculum, the time requirements may make assessment by GRS unsustainable. Similarly, an attempt at ongoing assessment of surgical residents’ intraoperative performance with a GRS would result in even more substantial time requirements, given that full surgical procedures commonly range from 2 to 8 hours in duration, substantially more than the approximately 5 minute duration of the individual technical tasks performed by the participants in the present work.

In addition to their time consuming nature, to reduce the bias and subjective nature of GRS, these assessment tools can be completed in a blinded fashion using video recordings. However to accomplish this, dedicated raters are needed to assess the videos, which is costly and requires significant infrastructure. Even if such raters were available, rater fatigue is problematic and can lead to decreased assessment quality. Rating videos daily for long periods of time is a monotonous and passive process. Intermittent assessment of inter-rater reliability to ensure calibration is maintained is important to maintain quality. However, again although doable in a
study setting for a relatively short and finite assessment period, using GRS as a routine measure of trainee competence in a surgical training program may not be a feasible solution. Therefore, seeking reliable objective assessments that are less time consuming, yet meaningful, is required to improve assessment of open technical skills and intraoperative procedural skills.

7.3.3 Adequate sampling

Studies within the field of surgical education are often limited by the sample size. Participants for these studies are recruited from within the medical system, specifically medical students, residents, fellows or faculty. This limits the potential number of participants available for recruitment. Furthermore, within the group of potential participants at a given training level, sub-populations exist that cannot always be considered similar for the purpose of many education research questions. For example, first year and fifth years residents are characterized by very different levels of technical skill, surgical judgment, decision making and patient management experience. Therefore, residents often cannot be studied as a single cohort.

Powering studies adequately has not been a problem for surgical education studies that compare two teaching paradigms or when constructing two technical skills curricula (Louridas, Bonrath, Sinclair, Dedy, & Grantcharov, 2015; Palter & Grantcharov, 2012). In chapter 6, a randomized controlled study was designed to assess the effect of MP on the interventional group compared to the conventional residency-training group. In this study, a sample size calculation of 10 participants per group was adequate to detect a difference in the primary outcome. However, as the surgical education field expands and studies attempt to adopt analytical techniques from the field of clinical epidemiology to explore the effect of technical skill on patient outcomes (Birkmeyer et al., 2013) or predictive modeling statistics to predict outcomes, larger sample sizes will be required. Sixty-five medical students were recruited for the technical skill-training curriculum described in chapter 5. This study is the largest of its kind in the surgical education literature and was able to identify performance clusters while exploring the relationship between clusters using non-parametric statistical tests. However, predictive modeling statistics or analytics would be very useful for this research. Creating learning curve models with adequate sensitivity and specificity to predict incoming trainees technical performance would require
hundreds of participants and therefore was not feasible for this study. Therefore, to overcome the inherent limitation of small sample sizes seen in surgical education, multicentre studies should be encouraged. Furthermore, national surgical education databases are not available to researchers in the field, further limiting the ability to aggregate national data over many years. Multicentre studies may also help with this limitation.
Chapter 8: Conclusion

Surgical programs endeavor to select medical students who are best suited to pursue a career in surgery. Once these students have been selected into the surgical training program the objective is to mentor these students over a five-year training period to graduate as safe independent surgeons. Of the many essential competencies acquired by training doctors, all surgical residents are required to reach technical competence prior to graduation. However, emerging evidence demonstrates that not all trainees reach this performance threshold (Cushchieri, 2003). Therefore, the research included in this thesis focused on seeking a reliable screening process for technical aptitude and assessed whether the cognitive training adjunct, MP, could be used to improve technical performance in surgical trainees.

The first study explored surrogate tests that may predict technical aptitude. It was hypothesized that surrogate markers, namely previous surgical experiences, non-surgical experiences and 2D-3D visual spatial tests correlate with baseline laparoscopic skills in the novice surgical trainee. However, of the surgical experience and non-surgical experiences studied, only previous laparoscopic experience predicted baseline surgical skill for incoming trainees. Similarly, visual spatial tests did not demonstrate a consistent association with overall technical performance. Thus, although the results were underwhelming, expecting surrogate markers to predict future performance may have been an over simplification of a complex process of many interacting abilities, rather than visual spatial ability alone.

Therefore, instead of relying on surrogate markers as predictors of technical skill, it was hypothesized that simulated technical tasks are better suited to stratify technical aptitude by assessing novice trainees’ learning curves over multiple repetitions. A national Delphi consensus questionnaire was distributed to program directors across Canada to gain their expert opinion on the selection of technical aptitude. PDs reported that approximately 5-15% for surgical trainees have difficulty reaching technical competence at the end of a standard five-year training program. This statement reconfirmed that continuing to seek a screening tool for technical aptitude was a worthwhile endeavor. The Delphi concluded that PDs reached positive consensus
that incorporating both basic open and laparoscopic tasks would be appropriate for incoming trainees. Therefore, open and laparoscopic technical skills were incorporated into the next study. The next study was designed to determine whether examining learning curves of simulated technical tasks over a one-month practice period would differentiate technical aptitude between students. It was hypothesized that medical students would display similar learning curves for disparate basic laparoscopic and open surgical skills, and that these will be correlated with their potential to reach proficiency in subsequent, more complex technical tasks. The results proved the hypothesis to be partly true. Four distinct learning curve clusters were identified. Cluster 1 selected students who demonstrated strong innate ability and were able to effortlessly learn all the technical skills, reaching competence quickly. Cluster 2 students were called high performers because they all eventually reached competence in all tasks but took slightly more time than their cluster 1 peers. Cluster 3 students were moderate performers and demonstrated difficulty, reaching proficiency in the more complex laparoscopic tasks. Cluster 4 individuals were unable to reach technical competence in 4 of 5 tasks including all laparoscopic tasks.

On examining the learning curves, it became apparent that cluster 1 and 2 had strong innate ability when compared to cluster 3 and 4 students. Furthermore, cluster 3 and 4 individuals did not catch up to their peers in cluster 1 and 2 over the training period, and demonstrated a continued disadvantage. However, when comparing the groups, cluster 3 individuals demonstrated somewhat stable performance below the competency threshold whereas cluster 4 individuals demonstrated large variability throughout their learning curves with no progression towards a plateau phase. Therefore, it was concluded from this study that learning curve clusters could be used as a screening tool to identify trainees with different levels of innate ability. Furthermore, screening for cluster 4 individuals should be considered due to their poor innate ability, inability to reach the proficiency threshold and variable unstable performance over the full durations of their learning curve.

In the final study, it was hypothesized that MP aimed at teaching the visual and kinesthetic cues for the crucial operative steps in a laparoscopic jejunojenuostomy (JJ), as well as for the management of adverse situations, improves surgical technical and non-technical performance and decreases stress levels experienced by the surgeon. Mental practice significantly improved advanced laparoscopic technical skill in the interventional group. Counter to the hypothesis,
however, MP was not associated with differences in stress levels or improvement in non-technical skills in the simulated crisis scenario.

Overall this thesis has proposed a selection curriculum from incoming trainees to assess technical skill ability by assessing learning curves and demonstrated the effectiveness of MP as an adjunct to technical performance in senior surgical trainees.
Chapter 9: Future directions

This thesis focused on selection and in-training assessment of technical skills, with the findings identifying a number of important directions for future research. Firstly, in the area of selection and technical skills, there is a need to explore alternative strategies to the use of GRS assessments, for assessment to be feasible in longitudinal studies to be able to measure predictive validity. Second, in additional to technical skill, future work should focus on improving current selection processes with the creation of predictive objective assessment tools for the other CanMEDs competencies outlined by the Royal College of Physicians and surgeons. Third, future research is needed to assess the effect of MP practice in the real operating room. Fourth, measuring whether students performed MP has only been accomplished using rating scale questionnaires, therefore more rigorous evaluations may be beneficial using functional imaging modalities.

9.1.1 Exploring intraoperative technical skills assessment

To establish predictive validity for selection clustering described in Chapter 5, it is essential to assess trainees objectively in the real operating room. However, further work is needed to identify assessment methods that can be efficiently and routinely used in this environment. Ideally, blinded GRS would be utilized with objective raters who are trained to use these assessment tools. However, as mentioned in the limitations section, this method is often not feasible due to the amount of manpower and the hours required to rate trainees’ performance. An alternate approach may be to train the faculty and nurses to use the GRS and rate the residents in real time. The difficulty for live assessments is obtaining timely scores. Having the operating room team complete the assessment directly after the case is often challenging given the other operational requirements involved in keeping the operating room running efficiently e.g. completing patient charting records, patient transfer, cleaning the room, preparation for the next case etc. In addition, bias is an inherent confound, because faculty and nurses work closely with the residents and therefore separating the personal dynamics to assess technical skill in isolation is difficult. However, preliminary evidence from the non-technical skills literature suggests that utilizing both surgical and non-surgical members of the operative team to assess surgical trainees may be worth investigating (Crossley, Marriott, Purdie, & Beard, 2011).
Another novel area of future research may be in video machine learning. This is a field gaining momentum in computer programming and data science. The concept that a computer can be programmed to interpret video and differentiate good and bad behaviors is fascinating and yet to be investigated in surgical education. Therefore, collaboration between video data scientists and surgeons may serve fruitful in working towards computer video assessments. If computers were able to reliably assess performance this would eliminate bias and human fatigue and promote sustainable intraoperative assessment.

Using these methods will allow students who have been assigned a performance cluster to be followed over time to assess whether their learning curve established during medical school is indicative of intraoperative performance during training.

9.1.2 Establish assessment measures of non-technical competencies

The work in this thesis focuses on technical performance. However, the technical performance of a surgeon is only one component of being a surgeon. Moreover, as demonstrated by the Delphi consensus questionnaire, various non-technical skills such as work ethic and passion for surgery were perceived as highly important by PDs. Furthermore, the Royal College of Physicians and Surgeons has clearly outlined seven roles in which each physician should strive to achieve competency during their training. These include the medical expert, communicator, collaborator, leader, health advocate, scholar and professional roles (["CanMEDS 2015: The next evolution of the CanMEDS Framework," 2013]). Investigating methods that assess attributes that are directly linked to these competencies is an essential step in all phases of training. Unfortunately, many of these attributes are difficult to measure using quantitative methodologies. Therefore qualitative or mixed method techniques may be more appropriate in understanding how to best assess these non-technical skills and how to incorporate them into our current selection process.

9.1.3 Quantifying mental practice objectively and transferring this techniques into the real operating room

Mental practice proved to be a successful adjunct to technical skills training within the simulated operating room even with added stress. Two areas of future research include translating this
technique into the real operating room, and quantifying MP using electroencephalograms (EEG) or functional magnetic resonance imaging (fMRI) for laparoscopic surgery.

The difficulty with studying the effect of MP on technical skill in the real operating room is isolating this intervention from the numerous confounders that may also affect technical performance. Firstly, each patient has anatomical differences that can make the same operation technically more or less challenging. Secondly, the operating room environment and personal that work within the OR may also change, which theoretically could also alter technical performance. Finally, fortunately adverse events are rare, and therefore measuring performance with respect to a direct stress is not feasible in the real operating room. However, there are validated metrics to adjust technical skill performance scores to account for case difficulty that could be incorporated during the assessment to control for some of these factors. Furthermore, selecting an advanced surgical technique that is very standardized, such as Roux-en-y bypass may also improve the feasibility of translating this technique into the operating room.

Another area of future research may be to measuring MP directly using functional neuroimaging modalities that record brain activity noninvasively, such as EEG or fMRI. Within laparoscopic surgery the use of MP has been quantified using validated questionnaires (e.g. MIQ), which are subjective in nature and therefore are prone to bias. However, authors have described using modalities such as EEG or fMRI to assess the use of mental imagery in real-time. The use of these modalities in the context of assessing surgeon performance during real or simulated surgical scenarios is in its infancy and has only been assessed in microsurgery but never in laparoscopy (Ros et al., 2009; Wanzel et al., 2007). Therefore, future studies may consider using these direct measures to assess whether mental imagery has improved between groups.
Copyright Acknowledgements

Chapter 2 (section 2.3.2)

Wolters Kluwer Health Lippincott Williams & Wilkins grants the authors permission to reuse their full-text article free of charge (non-commercial/ non-for-profit purposes) for the purpose of reuse in a dissertation.


Copyright (C) 2016 Wolters Kluwer Health, Inc. All rights reserved.

The final publication is available at:

Chapter 3

Copyright permission from Springer Science+Business Media


The final publication is available at:

Chapter 6

Copyright permission form the British Journal of Surgery through John Wiley and Sons states:

A. The following rules apply to all Users:
You may download, view, copy and save to hard disk or diskette and store or print out single copies of individual articles or items for your own personal and non-commercial use, scholarly, educational or scientific research or study. You may transmit to a third party, in hard copy or electronically, a single article or item from the Website for that party's own personal and non-commercial use or scholarly, educational, or scientific research or study. In addition, you have the right to use, with appropriate credit, brief excerpts from the Website content in your own work for the above purposes.


© 2014 BJS Society Ltd. Published by John Wiley & Sons Ltd

References

Accreditation Council for Graduate Medical Education. (2013). Work hour restrictions. December 4, 2013, from
http://www.acgme.org/acgmeweb/tabid/271/GraduateMedicalEducation/DutyHours.aspx

Accreditation Council for Graduate Medical Education. (2013). Accreditation Council for Graduate Medical Education, ACGME Program Requirements for Graduate Medical Education in General Surgery. Retrieved May 26, 2014, from
http://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/440_general_surgery_01012008_07012012.pdf


Admission requirements for Medical School at the University of Toronto. (2015). Retrieved Jan 22, 2015, from
http://www.md.utoronto.ca/admissions/information/requirements/Undergraduate.htm


Dreyfus, S. E., & Dreyfus, H. L. (1980). *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*.


Zevin, B., Levy Js Fau - Satava, R. M., Satava Rm Fau - Grantcharov, T. P., & Grantcharov, T. P. A consensus-based framework for design, validation, and implementation of simulation-based training curricula in surgery. (1879-1190 (Electronic)).