Towards Competency-Based Assessments in General Surgery

by

Peter Szasz

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

Institute of Medical Science
University of Toronto

Towards Competency-Based Assessments in General Surgery

Peter Szasz

Doctor of Philosophy

Institute of Medical Science
University of Toronto

2017

Abstract

Introduction: The training of surgical residents has seen a tremendous shift in education and assessment paradigms over the last decade. The traditional time-based framework, which has essentially remained unchanged since its conception over a hundred years ago is being replaced, albeit slowly, by a new competency-based framework. Despite the need for such change and this almost inevitable transition, there is currently a lack of evidence to support and operationalize many education and assessment facets within this new framework. The main goal of this thesis is to collect evidence that helps bridge many of the knowledge gaps currently present in order to support the transition towards competency-based training in General Surgery.

Methods: All data were prospectively collected and analyzed using descriptive and inferential statistics (both parametric and non-parametric). Research designs contained within the thesis include: a systematic review in accordance with current guidelines, a census survey, a consensus Delphi methodology and two cross-sectional observation studies. Traditional and contemporary frameworks of validity (and reliability) were adhered to throughout.

Results: Five research studies were completed. A systematic review, which documented the methods by which technical competence is currently assessed in surgical trainees, that there is a lack of performance standards in surgical education and how best to differentiate between the

terms competency and proficiency. An international education directorate's census survey, which acknowledged current and ideal assessment practices across various surgical jurisdictions and barriers to the implementation of competency-based assessments at all three phases of surgical training (selection, in-training and certification). A Delphi methodology, which created a new training and assessment model for varying levels of trainees in General Surgery, rooted in previously set guidelines. A cross-sectional study, which demonstrated that staff surgeons are reliable assessors of trainee technical and non-technical performance. Finally, a multi-center cross-sectional study, which set performance standards for technical and non-technical performance, demonstrated a discrepancy in some trainees' ability to meet both standards concurrently and examined factors that can predict standard (competence) acquisition.

Conclusions: The findings contained within this thesis contribute evidence to the operationalization of competency-based assessments in General Surgery.

Acknowledgments

First and foremost, I would like to thank Dr. Teodor Grantcharov for giving me the opportunity to be a part of his research group and for his continued and unwavering support, encouragement, backing, guidance, mentorship and friendship.

I would also like to thank my program advisory committee, Dr. Najma Ahmed and Dr. Charlotte Ringsted for their ongoing support, constructive feedback, guidance and advice. As well as Dr. Kenneth Harris from the Royal College of Physicians and Surgeons of Canada for his support of this work and for the great research relationship we have built over the years.

To my fellow research team members Dr. Marisa Louridas, Dr. Esther Bonrath, Dr. Nicolas Dedy, Dr. Sandra de Montbrun, Dr. Andras Fecso, Dr. Alaina Garbens, Dr. Mitchell Goldenberg and Mr. Karthik Raj. Your ongoing feedback, discussions over coffee and supportive research environment have been instrumental to my growth as a researcher and as a result the research contained in this thesis.

Thank you, Dr. Brett Howe, Dr. Michael Ott, Dr. Adam Fehr and Dr. Lloyd Mack, from Western University and the University of Calgary, respectively. Without your help prior to and during the performance standards study, my time at each University would have been fraught with logistical issues, and a lack of participant buy-in and faculty support.

Thank you, Dr. James Rutka in the Department of Surgery, Dr. Norman Rosenblum in the Clinician Investigator Program, Dr. Michael Fehlings from the Surgeon Scientist Training Program and the entire Division of General Surgery (especially Dr. Najma Ahmed, our program director) for giving me the opportunity and encouraging me to undertake this research as part of my residency training.

Finally, I could not be where I am today without the ongoing love and support of my wife, Cory, my parents Viera and Gabriel and my sister Sue. They have helped me more than I can describe and I am forever indebted to them and their steadfast encouragement and optimism. I have learned innumerous lessons from all of them and these have made me a better researcher, clinician and person.

Funding

This research work was funded by the H.S. Morton Fellowship Fund through the Royal College of Physicians and Surgeons of Canada, a Physicians' Services Incorporated (PSI) Foundation resident research grant, the John L. Provan Fellowship in Surgical Education from the University of Toronto, and a Ministry of Health – Clinician Investigator Program (MOH-CIP) award.

The Department of Surgery and Division of General Surgery also contributed funds (tuition and salary) in support of this research work.

Contributions

With the guidance of my supervisor and program advisory committee, I was the primary researcher involved and am responsible for all aspects of the work contained within the thesis. I am the first author on four of the manuscripts and a co-first author on one of the manuscripts contained in this thesis. My contributions include the conception, planning, design, data acquisition, data interpretation and drafting of each manuscript.

Dr. Teodor Grantcharov (supervisor) was involved in the conception, planning, design, data interpretation and critical review of each manuscript contained in Chapters 1,3,4,5,6. He also critically reviewed this thesis and provided mentorship.

Dr. Kenneth Harris was involved in the data interpretation and critical review of each manuscript contained in Chapters 1,3,4,5,6.

Dr. Najma Ahmed and Dr. Charlotte Ringsted (program advisory committee) were involved in providing thorough review and feedback for each study, as well as continued guidance and mentorship. They also critically reviewed this thesis.

Dr. Marisa Louridas was involved in the data interpretation and critical review of each manuscript contained in Chapters 1,4,5,6. In addition to the above she was also involved in the conception, planning, design and drafting of the manuscript in Chapter 3. Her major focus and contribution in Chapter 3 was on the selection time-point.

Dr. Sandra de Montbrun was involved in the data interpretation and critical review of each manuscript contained in Chapters 3 and 4.

Dr. Esther Bonrath was involved in the planning, data acquisition, data interpretation and critical review of the manuscript contained in Chapter 6.

Dr. Andras Fecso was involved in the data acquisition, data interpretation and critical review of the manuscript in Chapter 6.

Dr. Brett Howe, Dr. Adam Fehr, Dr. Michael Ott and Dr. Lloyd Mack, from Western University and the University of Calgary worked in helping the study in Chapter 6 succeed as a multi-intuitional project. They were also involved in the data interpretation and critical review of the manuscript in Chapter 6.

Mr. Anton Svendrovski aided in the statistical analysis for the manuscripts in Chapter 5 and 6.

Dr. Nicolas Dedy, Dr. David Wang, Dr. Sara Elkabany, Dr. Bijan Dastgheib and Dr. Bojan Macanovic were involved in the data acquisition (as raters) for the manuscript in Chapter 6.

Dr. Rajesh Aggarwal was involved in the data interpretation and critical review of the manuscript in Chapter 1.

Ms. Teruko Kishibe aided in the library search for the manuscript in Chapter 1.

Table of Contents

List of Abbreviations

ABMS	American Board of Medical Specialties
ABS	American Board of Surgery
ABSITE	American Board of Surgery In-Training Examination
ACGME	Accreditation Council for Graduate Medical Education
ARCP	Annual Review of Competence Progression
AUC	Area Under the Curve
BMAT	BioMedical Admission Test
CAGS	Canadian Association of General Surgeons
CBD	Case Based Discussions
CBME	Competency-Based Medical Education
CCT	Certificate of Completion of Training
CEX	Clinical Evaluation Exercise
CPSO	College of Physicians and Surgeons of Ontario
CQI	Continuing Quality Improvement
CT	Core Trainee
DOPS	Direct Observations of Procedural Skills in Surgery
DRIFT	Differential Rater Function over Time
ED	Education Directorate

EU	European Union
EWTD	European Working Times Directive
FLS	Fundamentals of Laparoscopic Surgery
FNA	Fine-Needle Aspiration
FOR	Frame of Reference
FPR	False Positive Rate
GAMSAT	Graduate Medical Schools Admissions Test
GMC	General Medical Council
GMP	Good Medical Practice
GRADE	Grading of Recommendations Assessment, Development and Evaluation
G Theory	Generalizability Theory
ICC	Intra-Class Correlation Coefficient
ICSAD	Imperial College Surgical Assessment Device
IQR	Inter-Quartile Range
IRT	Item Response Theory
ISCP	Intercollegiate Surgical Curriculum Programme
ITER	In-Training Evaluation Report
SJT	Situational Judgment Test
MCCQE	Medical Council of Canada Qualifying Examination
MEd	Master of Education

MESH	Medical Subject Headers
MFRM	Multi-Faceted Rasch Model
MIS	Minimally Invasive Surgery
MRCP	Member of the Royal College of Physicians
MRSC	Member of the Royal College of Surgeons
MSc	Master of Science
MSF	Multi-Source Feedback
M_uD	Mean of the Unsigned Difference
NBME	National Board of Medical Examiners
NOTECHs	Non-Technical Skills
NOTSS	Non-Technical Skills for Surgeons
OR	Operating Room
OSANTS	Objective Structured Assessment of Non-Technical Skills
OSATS	Objective Structured Assessment of Technical Skills
OSCE	Objective Structured Clinical Examination
OTAS	Observational Teamwork Assessment for Surgery
PBA	Procedure Based Assessments
PDT	Performance Dimension Training
PGME	Postgraduate Medical Education
PGY	Postgraduate Year

PhD	Doctor of Philosophy
PicSOR	Pictorial Surface Orientation
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCPSC	Royal College of Physicians and Surgeons of Canada
RCT	Randomized Controlled Trial
REB	Research Ethics Board
RET	Rater Error Training
RETREAT	Research, Education and Innovation for Better Outcomes
ROC	Receiver Operating Characteristic
SCORE	Surgical Council on Resident Education
SD	Standard Deviation
SDP	Scottish Doctor Project
SILS	Single Incision Laparoscopic Surgery
ST	Specialist Trainee
TNR	True Negative Rate
TPR	True Positive Rate
UGME	Undergraduate Medical Education
UK	United Kingdom
UKCAT	UK Clinical Aptitude Test
USMLE	United States Medical Licensing Examinations

List of Tables

List of Figures

List of Appendices

CHAPTER 1

INTRODUCTION

Chapter purpose

This chapter provides an understanding of the current education and assessment paradigms used in surgical education, the recent changes that have affected surgical training and how a transition towards competency-based medical education (CBME) and assessment may alleviate limitations of the current system in light of such changes. Current methods for technical and non-technical performance assessments are also discussed, as are pertinent topics related to the appropriate use of performance assessors and the incorporation of performance standards into surgical training.

Chapter preface

The contents of section 1.6.2 in this chapter, have been published in *Annals of Surgery* and reprinted with permission from Wolters Kluwer Health Lippincott Williams & Wilkins © as: Assessing Technical Competence in Surgical Trainees: A Systematic Review. **P. Szasz**, M. Louridas, K.A. Harris, R. Aggarwal, T.P. Grantcharov (2015) *Ann Surg* 261(6) 1046-1055 DOI: 10.1097/SLA.0000000000000866.

1 Introduction

The medical and surgical regulatory authority for resident training in Canada – the Royal College of Physicians and Surgeons of Canada (RCPSC) is in the process of implementing CBME with an emphasis on the attainment and demonstration of specific milestones, defined as " clear markers of expected resident performance at various stages in training, enabling educators to identify individual learner needs and abilities at an earlier stage" ("Competence by Design (CBD): Moving towards competency-based medical education," 2014; "Royal College of Physicians and Surgeons of Canada," 2014). Currently in Canada, the RCPSC examines trainees at the completion of training using an oral and multiple choice examination in order to determine if they are adequately trained to begin independent practice (from a judgment and knowledge standpoint)("Format of the Comprehensive Objective Examination in General Surgery," 2014; "Royal College of Physicians and Surgeons of Canada," 2014). Although, judgment and knowledge are also assessed during surgical training, technical performance (i.e. performing an operation or task) and non-technical performance (i.e. situational awareness, decision making, professionalism and communication etc.) are never assessed at the completion of training, nor more notably during surgical training where performance deficits if recognized can still be remedied ("Format of the Comprehensive Objective Examination in General Surgery," 2014; J. R. Frank, Snell, L.S., Sherbino, J, 2014).

The focus of this thesis will be to create defensible standards for both technical and non-technical performance that will allow the implementation of such performance assessments into residency training. These standards can then serve as milestones that need to be achieved in order for trainees to advance in their training, filling the void that currently exists. The subsequent work in this thesis collects evidence to support the transition to CBME. Prior to describing our work however, the historical surgical perspective and the current education and assessment landscapes need to be explained. As do the various methods by which technical and non-technical performance are assessed, the appropriate use of diverse assessors and the various performance standards that are currently employed in education. Where applicable in the thesis the Canadian training system is compared/contrasted with the United States and the United Kingdom (UK) training systems.

1.1 Current education paradigms

1.1.1 Historical perspective

Surgical training in North America as it is known today dates back 127 years, first introduced in 1889 by Dr. Halsted, the inaugural chief of the department of surgery at Johns Hopkins University (Halsted, 1904; O'Shea, 2008). Prior to his post as the chief of surgery, Dr. Halsted completed training in Austria and Germany, where his observations inspired him to implement a modified replica of the European system (O'Shea, 2008; Pellegrini, 2006). Up to that point the North American system was viewed as decades behind its European counterparts, consisting mainly of unregulated apprenticeships with a limited education structure (O'Shea, 2008; Polavarapu, Kulaylat, Sun, & Hamed, 2013).

Dr. Halsted initiated what became known as the pyramidal residency-training model (Halsted, 1904; Pellegrini, 2006). The main features of this model included intense competition among surgical trainees, graded surgical autonomy and the opportunity to work on and care for real patients (O'Shea, 2008; Polavarapu et al., 2013). Resident training occurred across a variety of surgical procedures in General Surgery under the tutelage of a senior surgeon, with the goal of creating a very small number of superiorly trained resident surgeons (O'Shea, 2008; Polavarapu et al., 2013). These resident surgeons would then go on to establish major surgical and research careers and populate the leading academic institutions within North America and abroad (O'Shea, 2008; Polavarapu et al., 2013). In the Halsted model, all entering surgical residents were given the opportunity to become the chief resident, the epitome of this pyramidal hierarchy (O'Shea, 2008; Polavarapu et al., 2013). However, only a small subset of these residents (often just one) would achieve such distinction yearly, with the remainder having no guarantee to become staff surgeons themselves (remaining assistants for prolonged periods of time) (O'Shea, 2008). The duration of training differed depending on the achievement of the resident and how high they were able to climb within this pyramidal training model, usually lasting between one and four and a half years (O'Shea, 2008). Although this model showed promise and was adapted by many institutions, it had multiple drawbacks (O'Shea, 2008; Pellegrini, 2006). These included the intense competition amongst trainees as a result of a hierarchical system, the power imbalance between the expert senior surgeon and the compliant resident surgeon, and the

creation of inadequately skilled surgeons (trainees that left the program after one year) (Grillo, 2004; Pellegrini, 2006).

Given these shortcomings to the Halsted model, Dr. Churchill proposed a modification, known as the rectangular residency-training model after his appointment in 1931 as a chair of surgery at Massachusetts General Hospital and the first residents matriculated in 1946 (Churchill, 1939; Grillo, 2004; Pellegrini, 2006). The main features of this model like the Halsted model included graded surgical autonomy and caring for real patients, but it focused more on a collaborative relationship between trainees to perpetuate their growth (Grillo, 2004). Resident training again occurred across a variety of disciplines, but now with an opportunity to gain additional skills through specialization in addition to General Surgery, under the tutelage of multiple staff surgeons and not just a single surgeon (Churchill, 1939; Grillo, 2004). The goal of the Churchill model was to ensure each resident surgeon was able and ready for independent practice and not to only support the most highly touted trainees (Grillo, 2004). As a result, all resident surgeons were selected for a five-year term with a selective few staying on for an additional two years to prepare them to become professors of surgery (in keeping with the primary goal of Halsted's model) (Churchill, 1939; Grillo, 2004). Furthermore, the Churchill model aimed to break down the boundaries between the senior and resident surgeons to create a less subordinate system that encouraged continual learning for both the resident and staff surgeons (Grillo, 2004).

The changes proposed and subsequently implemented by Churchill, coincided closely with the transition in Canada and the United States towards more structured and standardized surgical training, epitomized by the creation of the RCPSC and the American Board of Surgery (ABS) in 1929 and 1937, respectively ("The American Board of Surgery: About Us," 2015; Grillo, 2004; "Royal College of Physicians and Surgeons of Canada: History," 2014). Although there have been a few small changes since the model implemented by Churchill, residency training has remained essentially unaltered over the last 70 years.

1.1.2 Current surgical training pathway in Canada

Medical training in Canada is divided into undergraduate medical education (UGME) and postgraduate medical education (PGME) (Andrew, Oswald, & Stobart, 2014; "A collective vision for postgraduate medical education in Canada," 2012). UGME is the portion of the Canadian medical education system responsible for training medical students, while PGME is the portion responsible for training residents, all of whom have successfully completed medical school.

In Canada there are currently 17 medical schools ("Admission Requirements of Canadian Faculties of Medicine," 2015). Admission into the UGME program is based on several criteria, often, but not always including the completion of a four-year undergraduate degree with a broad focus on the physical sciences, social sciences and humanities ("Admission Requirements of Canadian Faculties of Medicine," 2015; Bandiera, Maniate, Hanson, Woods, & Hodges, 2015). Furthermore applicant's academic performance is reviewed, as are the results of the Medical College Admissions Test (MCAT), their extracurricular activities and performance during an interview (Bandiera et al., 2015). Once applicants have matriculated into a Canadian medical school they undergo either a four-year curriculum or in a select few medical schools a three-year curriculum, structured around the basic sciences, with the transition towards the clinical sciences and eventual clinical rotations through the common specialties (General Surgery, General Internal Medicine, Pediatrics, Psychiatry, Family Medicine etc.) learning on patients ("Admission Requirements of Canadian Faculties of Medicine," 2015). This curriculum culminates in the successful completion of all such rotations, a licensing examination - the Medical Council of Canada Qualifying Examination (MCCQE) part 1 and finally the transition into residency training via a national specialty specific trainee – program match ("Canadian Resident Matching Service (CaRMS) - examinations and assessments ", 2015).

The RCPSC oversees 79 different PGME programs dispersed among the hospitals of the 17 institutions (all PGME programs are associated with a medical school) ("Royal College of Physicians and Surgeons of Canada: History," 2014). These PGME programs can broadly be divided into specialties and subspecialties and each have a focus in either medicine (i.e. General Internal Medicine, Pediatrics etc.) or surgery (General Surgery, Orthopedic Surgery, Neurosurgery etc.). Specialty programs admit trainees directly from medical school, are usually

four to five years in duration and have broad scopes that are appropriate for trainees who wish to complete the program and enter independent practice or those that wish to complete more subspecialized training ("Royal College Discipline Recognition," 2014). Subspecialty programs admit trainees who have already completed a specialty program, are usually one to two years in duration and have more concentrated and advanced scopes ("Royal College Discipline Recognition," 2014). An example of a specialty program is General Surgery, while an example of a subspecialty program is Colorectal Surgery.

General Surgery itself is one of two original PGME specialty programs alongside General Internal Medicine introduced by the RCPSC upon its inception in 1929 ("The Future of General Surgery: Evolving to meet a changing practice," 2014; "Royal College of Physicians and Surgeons of Canada: History," 2014). At the time of the original PGME programs, the practice of surgery and General Surgery were synonymous. Over the last 80 plus years, surgery has been subdivided into multiple specialties (i.e. Neurosurgery, Otolaryngology, Orthopedic Surgery, Vascular Surgery, Cardiac Surgery, Urology, Plastic Surgery and Ophthalmology) with General Surgery being one of those specialties ("The Future of General Surgery: Evolving to meet a changing practice," 2014). The practice of a General Surgeon across Canada can vary depending on whether they are located in a major academic center, the community or in a rural setting ("General Surgery Profile," 2015). Regardless of geographic location, all General Surgeons provide care to adults, directed at the contents of the abdomen and pelvis, alimentary tract, breast, skin, soft tissue, and injuries as a result of trauma ("General Surgery Profile," 2015). Some General Surgeons based on extra-training they have undertaken (subspecialization as discussed above) and/or their geographic location may provide care in addition to General Surgery. This care may include Pediatric Surgery, Endocrine Surgery, Vascular Surgery, Surgical Oncology, Transplantation and Critical Care Medicine ("General Surgery Profile," 2015).

General Surgery specialty training in Canada is a five-year program ("Specialty Training Requirements in General Surgery," 2015). Trainees at certain institutions can decide to undertake research training during this time in the form of a Master of Science (MSc), Master of Education (MEd) or Doctor of Philosophy (PhD) and thereby extend their training anywhere from two to five additional years ("Special Program Training Requirements for the Clinician Investigator Program (CIP)," 2015; "Specialty Training Requirements in General Surgery," 2015).

The first two years of training are referred to as the junior or foundational years – undertaken by all trainees in any surgical specialty ("Objectives of Surgical Foundations Training," 2014). In General Surgery specifically, during these foundational years trainees spend at most 18 months rotating through different General Surgery blocks for one to three months at a time ("Objectives of Surgical Foundations Training," 2014; "Specialty Training Requirements in General Surgery," 2015). The remaining six months are spent on other surgical/medical services (i.e. Thoracic Surgery, Emergency Medicine, Anesthesia, Critical Care Medicine, Gastroenterology etc.) again for one to three months, in order to help round out the trainee so they may better understand and care for their patients ("Objectives of Surgical Foundations Training," 2014; "Specialty Training Requirements in General Surgery," 2015). At the completion of these two foundational years, trainees must successfully complete part 2 of the MCCQE and the RCPSC surgical foundations examination in order to progress into the more senior, specialty specific years ("Specialty Training Requirements in General Surgery," 2015).

The senior years are spent exclusively on General Surgery rotations or subspecialized General Surgery related services (i.e. Pediatric Surgery, Transplant Surgery, Surgical Oncology, Vascular Surgery etc.) ("Specialty Training Requirements in General Surgery," 2015). It is during the senior years that trainees acquire the knowledge, judgment and skill to operate (Drake, Horvath, Goldin, & Gow, 2013; "Objectives of Training in the Specialty of General Surgery," 2010). As trainees progress in their senior years, their autonomy increases in both the operating room and on the wards, especially for the more common patient presentations and procedures, as staff surgeon guidance and reinforcement decreases (Kempenich, Willis, Rakosi, Wiersch, & Schenarts, 2015; Soper & DaRosa, 2014). It is by this stepwise increase in autonomy that trainees eventually learn how to function and reason as independent surgeons (Aziz, 2015; Soper & DaRosa, 2014). At the completion of these three years, trainees become eligible to sit for the RCPSC licensing examination ("Specialty Training Requirements in General Surgery," 2015). RCPSC certification is then granted to those trainees who have completed five years of training in total, and have passed part 2 of the MCCQE and both the surgical foundations and certification examinations ("Specialty Training Requirements in General Surgery," 2015). After this point trainees can begin unsupervised independent practice in the field of General Surgery.

Regardless of whether trainees are junior or senior residents, their education is dependent on both formal and informal learning opportunities and experiences that together contribute to the acquisition of their knowledge, judgment and variety of skills.

The major method of formal learning is through a structured academic curriculum ("General Standards Applicable to All Residency Programs: B Standards," 2013; Sachdeva et al., 2007). Although the specifics of this may vary from program to program in Canada, the curriculum often takes place on a weekly basis for several hours in a centralized manner (General Surgery residents although dispersed among many hospitals in a given city learn together in one location). This curriculum covers the major basic and clinical topics in General Surgery through didactic lectures, interactive group discussions and hands on skills activities in the anatomy laboratory or simulation center ("General Standards Applicable to All Residency Programs: B Standards," 2013). In addition to this centralized curriculum, other formal learning platforms include more decentralized and hospital specific weekly education seminars, tumor discussion boards and continuing quality improvement (CQI) rounds where patient complications and mortalities are discussed ("General Standards Applicable to All Residency Programs: B Standards," 2013; Sachdeva et al., 2007). Furthermore, trainees alongside staff surgeons take part in journal clubs, where the newest evidence regarding a surgical topic is discussed, evaluated and debated (Crank-Patton, Fisher, & Toedter, 2001; Gatcliffe & Coleman, 2008; Gonzalo, Yang, & Huang, 2012; Shifflette, Mitchell, Mangram, & Dunn, 2012; Thompson & Prior, 1992). Finally, there are opportunities for trainees to attend education conferences and workshops geared at learning or improving specific areas of knowledge and/or surgical skill ("SAGES Advanced Laparoscopic Workshops for Surgical Residents," 2016).

Informal learning opportunities can be divided into workplace learning and self-directed learning (Blumberg, 1992; Monkhouse, 2010). In General Surgery, workplace learning is occurring all the time (Monkhouse, 2010). Examples include 1) morning rounds on patients admitted under the General Surgery service - where patient management plans are created and dispositions decided, 2) participation in the operating room (OR) where trainees acquire technical skills, procedural knowledge, judgment and communication, and 3) participation in the out-patient clinic setting and surgical ward where knowledge, judgment, communication, advocacy and professionalism skills are learned ("General Standards Applicable to All Residency Programs: B Standards," 2013). Finally, on-call activities are the pinnacle of informal

resident learning whereby trainees look after patients in a variety of settings more autonomously than at any other time of the day, while also gaining the above-mentioned experiences and skills (Scally, Sandhu, Magas, Gauger, & Minter, 2015). Examples of self-directed learning include reading around patient operative and clinic cases, reading basic and clinical textbooks and teaching more junior trainees, be it residents or medical students (Sherbino, Joshi, & Lin, 2015).

It is mostly in these informal learning opportunities, be it in the workplace or self-directed, that trainees acquire soft skills, better known as non-technical skills (Youngson & Flin, 2010). As it currently stands such non-technical skills training although seen as important, has not been formally incorporated into most surgical training programs and it is assumed that trainees will pick these skills up as they progress through training ("Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0," 2012; Dedy, Fecso, Szasz, Bonrath, & Grantcharov, 2015; Greig, Higham, & Vaux, 2015).

1.1.3 Current surgical training pathways internationally

United States

The UGME system in the United States mirrors that of Canada in terms of its structure, duration and preferred candidate qualities. Not unlike Canada, the United States UGME curriculum culminates with successful completion of all medical school rotations, and licensing examinations – the United States Medical Licensing Examination (USMLE) parts 1 and 2 and finally the transition into residency training via a national specialty specific trainee – program match ("About the USMLE ", 2016; Marsden, 2006).

PGME in the United States is again similar to that of Canada. The Accreditation Council on Graduate Medical Education (ACGME) oversees 130 specialty and subspecialty programs across approximately 700 institutions (PGME programs can be associated with an institution that has a medical school or with institutions/hospitals that do not) ("Accreditation Council for Graduate Medical Education: About," 2016). The specialty programs range in duration from three to six years and admit trainees from medical school into one of two pathways, categorical or preliminary (Marsden, 2006; Sarosi, Silver, Ben-David, & Behrns, 2014; Sullivan et al., 2013). Categorical trainees are those that have a guaranteed spot in a program for the full term of training (Sarosi et al., 2014; Sullivan et al., 2013). While preliminary trainees are guaranteed

only one or two years of basic training, after which they must secure categorical full time spots in order to complete their training (Sarosi et al., 2014; Sullivan et al., 2013). The preliminary spots are less desirable than the categorical spots, and fill the service demands of the individual programs without the need to offer and pay for a full term of training (Christein, Cook, Enger, & Farley, 2006; Sullivan et al., 2013). Like in Canada, subspecialty programs admit trainees who have completed a specialty program and are usually one to three years in duration ("ACS - years of postgraduate training ", 2016).

General Surgery specialty training in the Unites States, like Canada is also a five-year program, with similar opportunities for research during training that may lengthen this duration ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015). The program is again divided into junior years where part 3 of the USMLE examination is completed ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015). Unlike in Canada, there is no surgical foundations examination that precludes the transition into senior residency. The senior years, like in Canada are where most of the surgical skills, knowledge and judgment are acquired and fine tuned, again with increasing trainee autonomy (Drake et al., 2013). The main difference between the United States and Canada is the path to independent practice. In the United States, trainees who have successfully competed all three steps of the USMLE examination and all the requirements of a five-year residency program, can begin independent practice ("ABS Booklet of Information Surgery," 2015; "Specialty and Subspecialty Certificates," 2016). The ABS certification examination can be completed to distinguish oneself from your peers who are not board certified, but it is not a prerequisite for independent practice ("ABS Booklet of Information Surgery," 2015).

Analogous to Canada, the education of residents in the United States occurs by way of both formal and informal methods described above.

United Kingdom

In the UK admission into UGME is generally directly from secondary school although trainees holding undergraduate degrees can also apply ("Entry Requirements for UK medical schools," 2015). Admittance into medical school varies between the various institutions but is often dependent on previous academic performance and the completion of an aptitude test either the BioMedical Admission Test (BMAT), the UK Clinical Aptitude Test (UKCAT) or the

Graduate Medical Schools Admissions Test (GAMSAT) ("Entry Requirements for UK medical schools," 2015). The duration of training in the UK varies between four to six years and the UGME curriculum culminates with the completion of all medical school activities, the successful completion of the nationally administered Situational Judgment Test (SJT) and the transition into a foundations residency programme ("GMC role in education and training," 2016; Patterson, 2015).

PGME in the UK differs markedly when compared to Canada and the United States. The General Medical Council (GMC) oversees 61 specialty programs across a multitude of institutions (M. Jones, Carr, A., & Montgomery, J, 2010). All trainees complete the foundations programme, where they work on and rotate through various medical and surgical specialties to gain basic knowledge and skills for two years ("Intercollegiate Surgical Curriculum Overview," 2013). After the foundations programme, trainees become either core or specialist trainees (CT or ST) for an additional two years, where they again rotate on various specialties, but this time focusing more on either the surgical or medical specialty routes ("General Surgery," 2015; "Intercollegiate Surgical Curriculum Overview," 2013). At the completion of these early CT or ST years they complete the Member of the Royal College of Surgeons (MRSC) examination or the Member of the Royal College of Physicians (MRCP) examinations depending on their choice of further pursuing surgery or medicine respectively, and apply to specialty specific programs, such as General Surgery, Vascular Surgery, Urology etc. that vary in their duration anywhere between four to six years ("General Surgery," 2015)

General Surgery specialty training in the UK is six years in duration ("General Surgery," 2015). The first dedicated year of General Surgery is referred to as ST3, following the above-mentioned CT or ST years ("General Surgery," 2015). Trainees can decide to extend this training if they wish to pursue academically oriented surgical careers later on, by completing research or teaching endeavors leading to advanced degrees ("General Surgery," 2015). Similar to Canada and the United States it is during these more senior years (ST3 and greater) that trainees learn the specific knowledge, judgment and skill of the specialty and how to operate ("General Surgery," 2015). At the completion this specialty training, trainees apply for a Certificate of Completion of Training (CCT) ("General Surgery," 2015). It is after the completion of all three stages of training (foundation programme, core training and specialist training), the MRCS examination, the Fellow of the Royal College of Surgeons (FRCS) examination and having received the CCT

that trainees can act as consultant surgeons and begin independent practice ("General Surgery," 2015; "Intercollegiate Specialty Examination in General Surgery," 2015).

Again, analogous to Canada and the United States, the education of residents in the UK occurs by way of both formal and informal methods described above.

1.2 Changes affecting surgical training

1.2.1 A collection of changes

The above mentioned surgical education paradigms used in Canada and abroad have been successful in training competent surgeons since they were first introduced and later modified by Halsted and Churchill, respectively (Sachdeva et al., 2007). Over the last few decades however, multiple changes have been afoot which now threaten the ability of training programs to continue to produce competent residents using the traditional paradigms (Mattar et al., 2013; Soper & DaRosa, 2014). These changes include: public opinion, work hour restrictions, fiscal constraints to training residents, the use of new technology, the ageing population and the recent emphasis on surgeon to volume outcomes.

1.2.2 Public opinion

The trial and error of surgical training, and residents' learning from their mistakes accepted only a few decades ago, is no longer suitable in today's society due to public opinion (Brindley, Jones, Grantcharov, & de Gara, 2012).

Examples have emerged dating back to the 1980s that have caused the public to become increasingly concerned with residency training (Asch & Parker, 1988). Perhaps none are more written about than the incident surrounding an 18-year old girl named Libby Zion, who died in a New York City hospital in 1984 (Asch & Parker, 1988). Following her death, due to complications sustained from a lethal drug interaction, multiple inquiries were convened (Asch & Parker, 1988; Rosenbaum & Lamas, 2012). It was determined that her mortality secondary to this drug interaction likely resulted from her care being provided overnight, solely by resident physicians, who were fatigued and under supervised (Asch & Parker, 1988; Rosenbaum & Lamas, 2012). The ramifications as a result of this case were enormous (Fabricant, Dy, Dare, & Bostrom, 2013; Rosenbaum & Lamas, 2012). First, the public became aware of the role that

trainees played in their care, and second it led to changes in healthcare delivery in the state of New York initially, while also serving as a catalyst nationally and internationally with the 2003 ACGME resident duty hour restriction mandate (discussed below) (Fabricant et al., 2013).

Even when appropriate care is delivered, the public is becoming increasingly concerned with real patients being used for training purposes (Akhtar et al., 2015; Baker, Misra, Manimala, Kuy, & Gantt, 2013). A recent study documented that although the public supports the involvement of residents in their care, their support decreases as the complexity of the case increases and is lowest if the trainees actually perform the cases themselves as the primary surgeon (Kempenich et al., 2015). Furthermore, an older study revealed that patients perceive that resident involvement, specifically in the earlier years of training leads to worse overall patient care (Kim, Gates, & Lo, 1998).

Finally, the increased litigation against physicians and surgeons, including trainees, has made staff surgeons more concerned and less interested about including trainees in the care of their patients (Mello, Studdert, & Brennan, 2003).

As a result of all of these findings pertaining to the impact public opinion has had on trainees' contributions, there has been a transition away from learning directly on patients, towards more simulated training modalities (Champion & Gallagher, 2003; Kneebone et al., 2006). This has understandably resulted in a decrease in OR and other learning opportunities for trainees, and although simulated training has shown some benefits, nothing can ultimately replace the learning environment afforded by real patient encounters (Palter & Grantcharov, 2014).

1.2.3 Work hour restrictions

Over the last decade, the length of trainee shifts and total hours worked per week have come under scrutiny (A. B. Blum et al., 2010; Imrie, Frank, Ahmed, Gorman, & Harris, 2013). This is a direct result of incidents such as that of Libby Zion and literature suggesting that fatigue as a result of sustained work hours impairs attention, memory, and performance, and may result in increased medical errors (Barger et al., 2006; A. B. Blum et al., 2010; Gohar et al., 2009; Lockley et al., 2004). Supporting these notions is literature demonstrating that 24-hours of

wakefulness is comparable if not worse than a blood alcohol level of 0.05% when it comes to cognitive impairment (Falleti, Maruff, Collie, Darby, & McStephen, 2003).

As a result of this accumulating literature, multiple education bodies have implemented work hour restrictions in the United States and across Europe, with less formal regulations taking hold in Canada. The ACGME in the Unites States first created work hour standards in 2003 and modified them in 2011 ("Common Program Requirements ", 2016). These standards stipulate that a trainee may work no more than 80-hours per week, for a maximum duration of 16-hours at one time for postgraduate year (PGY) 1 residents and 24-hours at one time for trainees at the PGY2 level or higher ("Common Program Requirements ", 2016; Rosenbaum & Lamas, 2012). Initially created in 1993, the European Union (EU) led initiative entitled the European Working Time Directive (EWTD) was modified on multiple occasions and eventually rolled out for residents in 2009 ("Directive 2003/88/ES of the European Parliament and of the Council of 4 November 2003 concerning aspects of the organisation of working time," 2003; Goddard, 2010). These standards stipulate that a trainee may work no more than 48-hours per week, for a maximum duration of 13-hours at one time, for all levels of trainees ("Directive 2003/88/ES of the European Parliament and of the Council of 4 November 2003 concerning aspects of the organisation of working time," 2003). In Canada, the RCPSC created a resident duty hours committee, which produced national recommendations in 2013 ("National Steering Committee on Resident Duty Hours," 2013; Pattani, Wu, & Dhalla, 2014). These recommendations do not suggest a maximum cap on work hours per week, but do suggest that the duration of one shift cannot exceed 26-hours (16-hours in Quebec) and that call shifts that are spent fully in hospital cannot occur more often than one in four days ("National Steering Committee on Resident Duty Hours," 2013; Pattani et al., 2014).

The affects associated with the implementation of work hour restrictions on resident education have been summarized, initially by Fletcher *et al*. and more recently by Ahmed et *al*. (N. Ahmed et al., 2014; Fletcher et al., 2005). The review by Fletcher *et al*, demonstrated that the impact of such restrictions on resident education and involvement in operative cases was mixed, with some studies suggesting a benefit and others showing no clear benefit or decline (Fletcher et al., 2005). While, Ahmed *et al*. demonstrated that resident education measured both objectively (examination results, clinical performance) and subjectively (resident opinion) had worsened since the implementation of such restrictions (N. Ahmed et al., 2014).

Although research evaluating the impact of work hour restrictions is ongoing across multiple jurisdictions. It is highly suggestive based on the available, more recent evidence that restrictions have for the most part negatively impacted the education, case exposure and case experience of surgical residents over the last few years (N. Ahmed et al., 2014).

1.2.4 Fiscal constraints

Training surgical residents takes time. Resident education increases the time it takes for a staff surgeon or the collective training institution to complete daily activities (Bridges & Diamond, 1999; Papandria et al., 2012; Puram et al., 2015; Taravella, Davidson, Erlanger, Guiton, & Gregory, 2014; von Strauss Und Torney, Dell-Kuster, Mechera, Rosenthal, & Langer, 2012). Nowhere is this increase in time more pronounced than the OR. The incorporation of trainees into the OR increases the duration of the operation and as a result, operating costs (Bridges & Diamond, 1999; Papandria et al., 2012; Puram et al., 2015; Taravella et al., 2014). Although, residents have been involved in the OR since the time of Halsted, the more recent focus on cost cutting, lean management strategies and fiscal restraint beholding hospitals, has directly translated this into a training issue (Collar et al., 2012; Fine, 2009; Soper & DaRosa, 2014).

Bridges *et al.* in 1999 were the first group to document the increase in time and cost associated with training General Surgery residents at a single institution (Bridges & Diamond, 1999). They demonstrated that residents increased the time to complete an operation by an average of 12.6 minutes, resulting in a per resident cost of $47,970.00 USD over the duration of training (Bridges & Diamond, 1999). They subsequently extrapolated this to all General Surgery trainees in the United States in 1997 and showed this resulted in a $53 million USD increase (Bridges & Diamond, 1999). More recently, Papandria *et al*. for various General Surgery procedures and von Strauss Und Torney *et al.* for the laparoscopic cholecystectomy, both in multi-institution studies supported this increase in operative times, ranging between 12 and 24 minutes per procedure (Papandria et al., 2012; von Strauss Und Torney et al., 2012). Papandria *et al*. did not associate this directly to cost, while von Strauss Und Torney *et al*. suggested a €17.57 increased cost per minute or €492 per case, when trainees were involved (Papandria et al., 2012; von Strauss Und Torney et al., 2012). This increase in time and cost has also been supported in

other surgical specialties, including ophthalmology and otolaryngology (Puram et al., 2015; Taravella et al., 2014).

As a result of this emphatic focus on cost on the part of hospitals, there is less time for trainees to learn in the OR in particular, and trainees are often rushed to do so. This has also resulted in a decrease in OR exposure for current trainees in addition to the decrease in exposure as a result of increased public scrutiny, discussed above (Traynor, 2011).

1.2.5 New technology

The changes that have occurred in General Surgery over the last 25 years as a result of the introduction of new technology, are paralleled in the field of surgery really only by the introduction of inhalation anesthesia and aseptic techniques in the mid 19th century (Rutkow, 2012). Take for example the cholecystectomy. Up until the early 1990s, the surgery was exclusively done via an open surgical approach, with the first laparoscopic cholecystectomy, taking place in 1985 (C. A. Blum & Adams, 2011; Kavic, 1998). Through the introduction of laparoscopy and more recently, single incision laparoscopic surgery (SILS) and robotic surgery, the same procedure that only 25 years ago was done by a single approach, can today be done in four very different and challenging ways (Pfluke et al., 2011; Wren & Curet, 2011).

Understanding the impact that this technology explosion has had on resident surgical training is in its infancy (A. E. Park, Lee, T.H., 2011; Sirinek, Willis, & Schwesinger, 2016). Multiple studies have demonstrated however, that with the uptake of laparoscopy in particular, the resident operative experience has been weakened in two main ways (R. Chung, Pham, Wojtasik, Chari, & Chen, 2003; O'Bryan & Dutro, 2008; A. E. Park, Lee, T.H., 2011; Sirinek et al., 2016). Firstly, by decreasing trainee exposure to open surgeries, with the amount of open cases completed by residents steadily decreasing since the advent of laparoscopy, and secondly, by also failing to train residents adequately in laparoscopy (R. Chung et al., 2003; O'Bryan & Dutro, 2008; A. E. Park, Lee, T.H., 2011; Sirinek et al., 2016). Comparing actual resident laparoscopic case numbers to the Minimally Invasive Surgery (MIS) Fellowship Council's set out target required for competence, showed a large discrepancy, even for commonly performed procedures (A. Park, Kavic, Lee, & Heniford, 2007). Furthermore, transfer of training, the concept that skills learned in one area of surgery (open) can help in acquiring other skills

(laparoscopic) has not shown an association between these two approaches, and therefore training in both approaches is clearly warranted (Figert, Park, Witzke, & Schwartz, 2001).

The manner by which the increased use of laparoscopy specifically dilutes resident training is not completely clear; except of course that it has significantly increased the operative approaches trainees need to now learn during residency (open and laparoscopic). Other considerations include diluting the training of a resident if only indirectly given that with new technologies everyone including the staff surgeons and fellows become learners and ultimately compete or take away the residents learning opportunities (Lewis & Klingensmith, 2012). Furthermore, the increase in complexity of a case done laparoscopically (compared to its simplicity done in an open manner) has very likely transitioned some procedures from being done by residents, to now exclusively being done by staff surgeons.

Although SILS and robotic surgery have not really taken a hold in mainstream General Surgery and their impact on surgical training is currently undocumented, the incorporation of laparoscopy has created a real training dilemma from a breadth and complexity standpoint for both staff surgeons and trainees.

1.2.6 Ageing population

Given health care breakthroughs over the last few decades, patients are living longer and those over the age of 65 now makeup 16.1% of the Canadian population ("Population by sex and age group ", 2015). This prolonged life expectancy, has given these patients more time to acquire multiple medical morbidities (Karlamangla et al., 2007). As a result, the General Surgery patient today is older and sicker than the General Surgery patient of the past, resulting in more complex operative interventions in both the elective and emergent settings, and in the post-operative care provided (Pofahl & Pories, 2003).

Studies completed over the last three decades evaluating a variety of General Surgical procedures have documented that older patients (≥ 65 years of age, but for the most part ≥ 80) have significant mortality rates ranging between 8% and 15.2% and morbidity rates ranging between 32% and 71.8% (Bufalari et al., 1996; Lees, Merani, Tauh, & Khadaroo, 2015; Racz, Dubois, Katchky, & Wall, 2012; Rigberg, Cole, Hiyama, & McFadden, 2000; Rorbaek-Madsen et al., 1992). The major predictors of complications did not relate to age per se, but rather the

number of underlying medical comorbidities and/or the overall fitness of patients undergoing surgery (Bufalari et al., 1996; Lees et al., 2015; Racz et al., 2012; Rigberg et al., 2000; Rorbaek-Madsen et al., 1992). In a review, Pofahl *et al.* corroborated such previous findings, for both elective and emergent surgeries, demonstrating higher morbidity and mortality rates in older patients compared to younger patients undergoing the same procedures (Pofahl & Pories, 2003). Reasons for these increased rates were attributed to differences in older patient physiology, the interplay and role of their underlying medical issues, altered post-operative care needs and perhaps most importantly in the context of this thesis, a lack of familiarity and experience (lack of training) among both staff surgeons and residents relating to geriatric surgery (Pofahl & Pories, 2003).

As a result, the complexities associated with operating on older patients with multiple comorbidities, the growing elderly population, and decreased trainee operative exposures in general, collectively inundate a training system already bursting at the seams.

1.2.7 Surgeon volume-outcome data

Recently in surgery there has been a significant interest and focus on operative case volumes (both at the hospital and individual surgeon levels) and patient outcomes (M. F. Brennan & Debas, 2004). This is the result of an initial study published in 1995 demonstrating that patients who underwent pancreatic resections for cancer in high volume centers (>81 cases/year) did much better from a morbidity and mortality perspective, than those who had surgery in low volume centers (< 50 cases/year) (Lieberman, Kilburn, Lindsey, & Brennan, 1995).

Since that initially study, this concept of case volume and patient outcomes has taken on a life of its own and spurred a significant amount of research focusing on all aspects of General Surgical practice (Al-Qurayshi, Robins, Hauch, Randolph, & Kandil, 2016; Aquina et al., 2015; Birkmeyer et al., 2002; Jeong, Ryu, Choi, Piao, & Park, 2014; Ricci et al., 2014; Skinner, Helsper, Deapen, Ye, & Sposto, 2003).

Specifically, Birkmeyer *et al*. reviewed hospital volumes and patient outcomes, demonstrating that for procedures such as the colectomy, gastrectomy, esophagectomy and pancreatic resection, low volume centers (defined differently for each procedure) carried a

substantially different and significantly higher rate of operative mortality than high volume centers (also defined differently for each procedure), between 7.4% - 23.1% and 3.8% - 8.7%, respectively (Birkmeyer et al., 2002). At the individual surgeon level, case experience has been shown to impact patient outcomes for both simple (breast surgery, inguinal hernia repair, thyroid) and complex (pancreatic resection, gastrectomy) procedures. Skinner *et al*. showed that for patients undergoing breast surgery for cancer, those who were operated on by surgical oncologists (breast specialists) with higher volumes per year (>15 cases/year typically), had a statistically better overall survival measured at five years than those undergoing surgery by non surgical oncologists with lower volumes (Skinner et al., 2003). For inguinal hernia repairs high volume surgeons (> 25 cases/year) had lower reoperation rates for hernia recurrence than low volume surgeons (<25 cases/year) (Aquina et al., 2015). Similarly for thyroidectomies for benign or malignant disease high volume surgeons (> 30 cases/year) had improved clinical outcomes compared to intermediate (4-29 cases/year) and low volume surgeons (1-3 cases/year) (Al-Qurayshi et al., 2016). For more complex procedures, Ricci *et al.* for a laparoscopic distal pancreatic resection, described the learning curve of a General Surgeon at which point their risk of operative morbidity decreased to be 17 cases, while Jeong *et al*. for a laparoscopic total gastrectomy found the learning curve of a General Surgeon at which point operative mortality decreased to be 45 cases (Jeong et al., 2014; Ricci et al., 2014).

It is evident that case volume plays a large role in improving patient outcomes. The influence of this recent focus on hospital and surgeon volume and the impact it has on resident training are not yet known. Conceivably however, as a result of these studies some procedures are moving away from low volume centers, where some residents are training, impacting their overall exposure. Furthermore, and perhaps more problematic is that in low volume centers where staff surgeons have limited experience, how can they be training residents properly to undertake these procedures, that they themselves only do rarely and may not be completely comfortable with?

1.2.8 Ultimate consequence of such changes

Over the last few decades there has been a prevailing opinion that graduating trainees are less prepared to undertake independent practice than their predecessors (Soper & DaRosa, 2014). The reasons stated point directly to these recent changes that have affected surgical education

and training in General Surgery, perhaps now more than ever before (Kempenich et al., 2015). As a result, solutions must be sought that will maintain the high standards needed for training residents and many of these solutions are already in progress such as the move towards CBME (as depicted in the sections to follow) (J. R. Frank, Snell, L.S., Sherbino, J, 2014). One specific area unrepresented at this current juncture, particularly important given these changes and the transition to CBME, relates specifically to what tasks and procedures should still be in the scope of training for General Surgery residents and what procedures should all residents be assessed on to delineate their competency, as we move through the 21st century. These two voids go on to form Aim # 3 in this thesis.

1.3 Competency-based medical education paradigms

1.3.1 Transition to competency-based medical education

The fact that trainees are operating on older, sicker patients, with less OR exposure as result of work hour restrictions, and a multitude of techniques at their disposal all under the keen eyes of governing bodies and the public, it's a wonder the current training system has survived this long. As a result of the above-described changes, the current education paradigms are no longer meeting resident educational needs and CBME has been slowly making its way into surgical training (J. R. Frank, Snell, L.S., Sherbino, J, 2014).

1.3.2 Basis of competency-based medical education

CBME is an outcome focused education (OFE) paradigm that is altering the way in which residency programs train their residents (J. R. Frank, Snell, L.S., Sherbino, J, 2014). CBME focuses less on the duration of training and more on the acquisition, maintenance, and demonstration of specific competencies (J. R. Frank, Snell, L.S., Sherbino, J, 2014).

1.3.3 Roots of competency-based medical education

CBME (really OFE) was first described approximately forty years ago (J. R. Frank, Snell, et al., 2010; Harden, 1999; Rubin, 1984). It is in direct contrast to traditional education models - where their main focus is on educational processes (J. R. Frank, Snell, et al., 2010; Harden, 1999). Instead in CBME (OFE), trainee and program outcomes supersede the educational process and guide all future curricular transformations (J. R. Frank, Snell, et al., 2010; Harden,

1999). Prior to its incorporation into PGME, it had successfully been implemented in various education fields including elementary and secondary school education, teacher education, and the military (J. R. Frank, Snell, et al., 2010; Klamen, Williams, Roberts, & Cianciolo, 2016).

1.3.4 Education theory underpinning competency-based medical education

CBME has its roots in behavior learning theory (Gonczi, 1997). The premise behind this theory is that all learning is based on observable behaviors that individuals encounter in the educational environment (Gonczi, 1997; Magnusson, 1990). Specifically, behavior learning theory states there is always a stimulus/response connection to behavior and that new behavior is learned through conditioning: either classical or operant (McLeod, 2013). Classical conditioning is the ability of trainees to associate two stimuli to produce a new behavior, while operant conditioning is based on the notion that all new behaviors are met with external responses that are either positive, negative or neutral (McLeod, 2014, 2015). Those behaviors that have positive external responses are prone to be repeated, those that are met with negative responses are prone to be discontinued and those that are met with neutral responses may or may not be repeated (McLeod, 2015).

In addition to behavior learning theory, Miller's pyramid of learning and assessment drives CBME (J. D. Beard, 2008; Crossley & Jolly, 2012; Miller, 1990; Shalhoub, Vesey, & Fitzgerald, 2014). Traditional surgical education and assessment are predicated on testing trainee knowledge or judgment, which make up the lowest levels of Miller's pyramid ('knows' and 'knows how') (Crossley & Jolly, 2012; Miller, 1990). CBME on the other hand, which requires the direct observation of trainee performance in the workplace (as discussed below) is associated with the higher levels of this pyramid, including 'shows how' or in certain circumstances even 'does' (J. D. Beard, 2008; Crossley & Jolly, 2012; Miller, 1990; Shalhoub et al., 2014).

1.3.5 Principles of competency-based medical education

There are four main principles that underlie CBME in PGME and these include: 1) a focus on outcomes, 2) the demonstration of multiple trainee abilities/competencies, 3) the decreased significance of training duration, and 4) the importance of individual trainee flexibility (J. R. Frank, Snell, et al., 2010).

A trainee demonstrating what they have learned forms the basis of CBME (C. Carraccio, Wolfsthal, Englander, Ferentz, & Martin, 2002; J. R. Frank, Snell, et al., 2010; Voorhees, 2001). The traditional presumption in medicine and surgery that sound and structured educational processes within a residency program will undoubtedly create physicians and surgeons that are ready for independent practice, no longer holds true as a result of the many changes affecting surgical education, described above (C. Carraccio et al., 2015). In CBME, it is the residency-training program and governing bodies' (i.e. RCPSC) responsibility that trainees continually have an opportunity to demonstrate and apply what they have learned and that this then forms the metric which determines whether they are ready and able to practice independently (C. Carraccio et al., 2002; J. R. Frank, Snell, et al., 2010).

Traditional medical education models are structured around training objectives, which are often extensive and non-specific ("Curriculum outline for General Surgery Residency," 2015-2016; J. R. Frank, Snell, et al., 2010). These objectives result in training programs and trainees focusing almost exclusively on medical knowledge at the expense of other abilities also important for medical practice (collaboration, communication, professionalism etc.) and in the compartmentalization of learned information into independent silos instead of a knowledge collective (C. Carraccio et al., 2015; J. R. Frank, Snell, et al., 2010). Instead, CBME focuses on multiple abilities/competencies (defined below in section 1.3.6), not just medical knowledge that can be built upon over time to create more holistic practitioners (C. Carraccio et al., 2015).

In traditional medical education models, time was by most accounts the main and only surrogate for competence (Hodges, 2010). In General Surgery, the training process is five years in duration and at the completion of those five years it was expected, trainees were ready to begin independent practice (Hodges, 2010). In CBME, it is the acquisition and demonstration of specific abilities/competencies important to independent practice that is paramount to the completion of training, notwithstanding the exact duration of that training (J. R. Frank, Snell, et al., 2010).

In keeping with a decreased focus on duration of training, the idea of learner centeredness has also emerged – the concept of having trainees take responsibility for their education alongside their training program (C. Carraccio et al., 2015; J. R. Frank, Mungroo, et al., 2010; Hodges, 2010). As CBME measures competence based on what trainees have actually acquired,

trainees will achieve competence in various domains at different rates (C. Carraccio et al., 2015). Instead of General Surgery training always being five years in duration, some trainees may be able to achieve all required competencies in four years, while others will require five years or even six years. CBME gives the training program and trainee the flexibility to grow and learn at their own pace, creating their own learning trajectories rather than trying to fit the same model onto each trainee (C. Carraccio et al., 2015; Schumacher, Englander, & Carraccio, 2013).

1.3.6 Domains of competency-based medical education

Within CBME, there are many domains that must be acquired, not just medical knowledge. These various domains deal with the various aspects that create a well rounded trainee and eventually independent practitioner, with the appropriate knowledge and judgment in regards to their chosen area of practice, as well as the ability to work within the complex healthcare system, all while providing care to patients (C. Carraccio et al., 2015).

These domains are termed competencies. The specifics of which vary based on the jurisdiction, however they share many commonalities and can broadly be categorized into: medical knowledge, technical performance, scholarship, collaboration, patient advocacy/healthcare management and leadership, communication, and professionalism ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; Cogbill, 2014; J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015; "The Good medical practice framework for appraisal and validation," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012).

1.3.7 Implementation into Canadian surgical training

In Canada, CBME is concentrated on seven domains or competencies and these include: medical expertise, communication, collaboration, leadership, health advocacy, scholarship, and professionalism (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015). Although all seven competencies are imperative, the system is centered on medical expertise, the central role of physicians in the care of patients (Figure 1) (J. R. Frank, Snell. L., Sherbino, J, 2015). The competencies/domains that form the basis of CBME were first developed in 1996 as the CanMEDS framework, an initiative created to improve the overall training of specialist physicians ("CanMEDS history," 2015). This framework was updated and fully adopted by the

RCPSC in 2005 forming the current training standards in each specialty ("CanMEDS history," 2015). In 2015, the framework was updated further to better align with the needs of CBME from both an education and assessment standpoint with the implementation of training milestones (described below in section. 1.5.2) ("CanMEDS history," 2015).

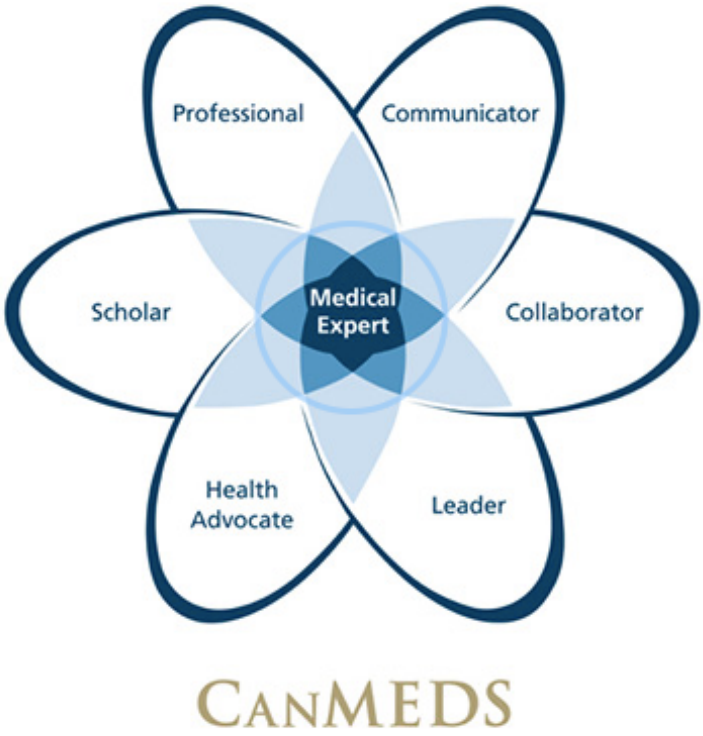

Figure 1. CanMEDS 2015 competency framework**.** Copyright © 2015 The Royal College of Physicians and Surgeons of Canada http://rcpsc.medical.org/canmeds. Reproduced with permission.

All specialty and subspecialty programs in PGME follow this seven competency model to training (J. R. Frank, Snell, L.S., Sherbino, J, 2014). Under each of the seven competences, there are more generic overall aims that apply to all residency programs regardless of focus, referred to as key competencies and more specific program aims, referred to as enabling competencies (J. R. Frank, Snell. L., Sherbino, J, 2015; "Objectives of Training in the Specialty of General Surgery," 2010). The implementation of the full CBME framework into PGME is underway and proceeding slowly ("Implementing CanMEDS 2015," 2015). Different specialty and subspecialty cohorts are being transitioned at various rates even amongst different institutions, to ensure a

compromise to resident education does not occur ("An iterative and adaptive approach to CBD implementation," 2015). The first cohort will include Otolaryngology and Medical Oncology in 2016 ("An iterative and adaptive approach to CBD implementation," 2015). Regardless of whether the full CBME framework has been implemented, the majority of training disciplines have embedded at least some CBME aspects into training - specifically at the medical expertise, scholarship, professionalism and communication domains. While a few programs in a self-initiated fashion, with the support from the RCPSC, have completely transitioned into CBME, namely the Orthopedic Surgery program at the University of Toronto (Alman, Ferguson, Kraemer, Nousiainen, & Reznick, 2013).

Starting in 2009, the Orthopedic Surgery program has divided their matriculating trainees into a traditional or CBME curriculum (Ferguson et al., 2013; N. Smith, Harnett, & Furey, 2015). Those in the CBME stream undergo intensive modular training focused on specific education objectives directed at all seven CanMEDS competencies with frequent and various assessments (Ferguson et al., 2013). When trainees have successfully completed a module by demonstrating appropriate performance across multiple competencies, they move onto the next module (Ferguson et al., 2013). In total there are 21 modules distributed along three phases of training (introductory, advanced and intermediate) (Ferguson et al., 2013). The Initial three year data have shown an improved performance across all assessment domains (in the operating room, on written examinations, on oral examinations and via allied health feedback on communication, teamwork and professionalism) for trainees in the CBME stream with some trainees being able to complete the training program in the shorter timeframe of four years (one year shorter than is the traditional RCPSC duration) (Ferguson et al., 2013). The findings also point to an acceleration in skill development and less wasted time on the part of the trainees doing non-education tasks, in the CBME stream (Ferguson et al., 2013). Follow up data evaluating the first cohort to graduation, are currently pending (Ferguson et al., 2013).

In General Surgery the transition towards full CBME will take place at the earliest as part of the third cohort, with a time to full implementation by 2022 ("An iterative and adaptive approach to CBD implementation," 2015).

1.3.8 Implementation into international surgical training

United States

In the United States, CBME is concentrated on six domains or competencies and these include: medical knowledge, patient care, practice-based learning and improvement, interpersonal and communication skills, systems-based practice, and professionalism ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; Holmboe et al., 2015). The ACGME in coordination with the American Board of Medical Specialties (ABMS) formed the six domains that embody CBME in 1999 (Holmboe et al., 2015; Nasca, Philibert, Brigham, & Flynn, 2012). The first specialties to create specific education indicators (subheadings under each of the six competencies) began in 2007 with Internal Medicine, Pediatrics and General Surgery (Holmboe et al., 2015; Swing et al., 2013). Since then all specialties have created such indicators, and for the first time in 2013 seven specialties including three surgical specialties (Urology, Orthopedic Surgery and Neurosurgery) started incorporating them into their training programs (Holmboe et al., 2015; Swing et al., 2013). This slow transition will occur over the next several years starting in 2014-2015, with programs reporting their outcomes to the ACGME (Holmboe et al., 2015). Although very early, literature coming out from a few of the early adopting specialties (2013) suggests that the incorporation of CBME has been mostly positive from a qualitative standpoint for improving resident education from both the trainee and program perspective, with ongoing work needed in areas of program operationalization (Aagaard et al., 2013; Angus, Moriarty, Nardino, Chmielewski, & Rosenblum, 2015). There is currently a paucity of literature directly evaluating the quantitative impact on the education of residents, with just one study showing that following the incorporation of CBME the ability of staff physicians to discriminate between different levels of trainee performances improved (Bartlett et al., 2015).

In General Surgery specifically, the transition to CBME in the United States, like in Canada, will occur over the next several years.

United Kingdom

In the UK, CBME is predicated on four domains and these include: knowledge/skills/performance, safety and quality, communication, partnership/teamwork, and

maintaining trust ("Good Medical Practice," 2014; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). The transition towards CBME in the UK began in the early 2000s spurred by both the Scottish Doctor Project (SDP) and shortly thereafter the GMC as part of their Good Medical Practice (GMP) guideline efforts (Simpson et al., 2002; "Tomorrow's Doctors," 2009). As opposed to Canada and the United States, the UK system went through multiple changes to arrive at their current four domain structure as part of their reworking of the GMP guidelines, which came into effect in 2013 ("Good Surgical Practice," 2014; "Tomorrow's Doctors," 2009; "The Trainee Doctor," 2011). Despite a multitude of reports documenting these changes and the need for all specialties to have been granted approval for having sufficiently incorporated aspects of the GMP guidelines into their curricula by 2010, there is a paucity of literature describing the implementation of CBME within the various specialties ("Good Medical Practice," 2014). This has led some in the UK system to suggest that the GMP guidelines are less about education and more about program regulation (Gray & Grant, 2012). Nonetheless, a recent report published in 2014 has molded the GMP guidelines specifically to surgical training in the UK, which may aid in its operationalization ("Good Surgical Practice," 2014).

The full transition into CBME within General Surgery will likely incorporate this recent report and occur over the next several years in coordination with the organization responsible for surgical training in the UK – the Intercollegiate Surgical Curriculum Programme (ISCP) (Gray & Grant, 2012; "Intercollegiate Surgical Curriculum Overview," 2013).

1.4 Current assessment paradigms

1.4.1 Differentiating between formative and summative assessments

In education there are two main types of assessment, formative and summative, each serving a distinct purpose (Epstein, 2007; Konopasek, Norcini, & Krupat, 2016). Formative assessments alternatively called low-stakes assessments are completed as a method of feedback to improve trainee-learning, understanding and to guide further education practices (Epstein, 2007; Hawkins et al., 2015; Holmboe, Sherbino, Long, Swing, & Frank, 2010; Konopasek et al., 2016). Formative assessments are completed on a regular basis by trained educators (staff physicians) who continuously work with the trainees (Epstein, 2007; Holmboe et al., 2010). Examples of formative assessments in PGME include: feedback given after the discussion of a

trainee with a patient, or feedback given at the completion of a surgical procedure. Conversely, summative assessments alternatively called high-stakes assessments are completed as method to determine what a trainee has learned up to that point and to decide whether the trainee can progress in their training or if they require a remediation curriculum (Epstein, 2007; Holmboe et al., 2010; Konopasek et al., 2016). Summative assessments are completed sparingly, often at the end of a prolonged rotation, year of training or at the completion of an entire residency (i.e. certification) using various methods, but often via structured examinations (Epstein, 2007; Hawkins et al., 2015; Holmboe et al., 2010). Examples of summative assessments in PGME include: rotation specific oral examinations, yearly written examinations and a certification examination at the completion of training.

1.4.2 Current assessment perspectives in Canada

Currently in Canada, PGME is structured around both formative and summative assessment modalities (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Hodges, 2010). The specific formative assessments differ depending on the type of specialty, and the duration of the residency program. The summative assessments also differ depending on these variables, but all PGME programs in Canada including General Surgery, culminate as discussed in section 1.1.2 with a certification examination ("Specialty Training Requirements in General Surgery," 2015). This certification examination is composed of a written and/or multiple-choice examination along with either an oral examination or an Objective Structured Clinical Examination (OSCE) where standardized patient encounters, clinical or practical components specific to each specialty are assessed ("Policies and Procedures for Certification and Fellowship," 2014). In some specialties a combination of an oral examination and OSCE are used ("Policies and Procedures for Certification and Fellowship," 2014).

In General Surgery specifically, formative assessments are composed of feedback from the surgical staff a trainee works with (M. R. Cook et al., 2015; Epstein, 2007). The opportunities for such formative assessments occur in didactic lectures, surgical rounds, the outpatient clinic, in the operating room and after a night on call, whereby the trainees' knowledge, judgment and when applicable technical skill are evaluated and options for improvement are discussed (M. R. Cook et al., 2015; Phillips, Madhavan, Bookless, & Macafee, 2015). The major issues with the current formative assessment paradigms however are that they don't happen as often as they

should, and although there are multiple opportunities, such feedback realistically occurs quite rarely, perhaps only on a few occasions during an entire surgical rotation (Hoffman, Petrosky, Eskander, Selby, & Kulaylat, 2015; Jensen, Wright, Kim, Horvath, & Calhoun, 2012). Furthermore, although formative in nature and not high-stakes, this does not mean that they are unimportant or insignificant. Yet the staff surgeons who provide this feedback are often untrained in doing so and they themselves receive little real guidance on their ability to provide quality formative assessments (M. Ahmed, Sevdalis, Vincent, & Arora, 2013; Hoffman et al., 2015; J. Norcini et al., 2011). Finally, the other competencies/domains within the CBME system outside of knowledge, judgment and technical skill (all of these falling under medical expertise) are rarely discussed (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015). With a lack of assessor training and the omission of most competencies/domains, how useful and instructive can such formative assessments really be? It seems that in the current Canadian General Surgery training system, formative assessments have many avenues for improvement.

In contrast, summative assessments are seemingly done better in PGME (including General Surgery training) currently, likely as a result of their structured nature, often taking the form of a recognized examination (Dannefer, 2013; Epstein, 2007). The current summative assessment methods utilized include end of rotation structured debriefing sessions, where in-training evaluation reports (ITERs) are completed by the education representative of that particular hospital and the trainee, with how the trainee fared during their time on that rotation (Compeau, Tyrwhitt, Shargall, & Rotstein, 2009; Dudek & Dojeiji, 2014). These ITERs often assess trainee knowledge, judgment, and skill, and can also theoretically evaluate non-technical performance such as teamwork, professionalism and communication (Kassam, Donnon, & Rigby, 2014). Other methods for summative assessment vary based on the specific institution training General Surgery residents but often includes a yearly in-training Canadian Association of General Surgeons (CAGS) examination taken by all levels of residents and a yearly oral examination (Maciver, 2016). Finally, the RCPSC certification examination is the penultimate summative assessment which in General Surgery is composed of a multiple choice examination which tests knowledge and an oral examination which tests judgment ("Specialty Training Requirements in General Surgery," 2015). Like their formative assessment counterparts in General Surgery, summative assessments also currently have a few shortcomings. Firstly, the ITER is often completed weeks to months after a trainee has finished a rotation, often providing

inaccurate (lack of valid and reliable) assessments of a trainee due to recall bias, and although it theoretically assesses non-technical performance, this is completed in a subjective manner by way of asking staff surgeons or other team members the trainee may have worked with, who aren't trained and often don't understand many of the non-technical constructs actually under evaluation (Dedy, Szasz, et al., 2015; Ginsburg, Eva, & Regehr, 2013; Kwolek et al., 1997; Patel, 2015; Youngson & Flin, 2010). There is never a summative assessment of non-technical performance by anything other than the ITER – where trained assessors are not utilized ("Specialty Training Requirements in General Surgery," 2015). Secondly, at no time in General Surgery training is technical performance actually assessed in a summative manner ("Specialty Training Requirements in General Surgery," 2015).

1.4.3 Current assessment perspectives internationally

United States

In the United States, PGME is again structured around both formative and summative assessments, the particulars of which vary depending on the specialty and duration of program ("Common Program Requirements ", 2016). Unlike in Canada however, as discussed in section. 1.1.3 there is no mandatory summative assessment in the form of a credentialing examination that must be completed in order to begin independent practice ("Specialty and Subspecialty Certificates," 2016). In the United States if undertaken, the credentialing examinations are composed of both written, oral and OSCE components that may be completed at various times starting in residency and into the first several years of practice ("Specialty and Subspecialty Certificates," 2016). These examinations are not mandatory but can be completed to make a trainee or practicing physician stand out compared to their colleagues ("Specialty and Subspecialty Certificates," 2016)

In General Surgery specifically, there are opportunities for residents to be assessed in a formative manner and be provided with feedback after any educational activity, just like in Canada. The United States like Canada however, suffers from the same shortcomings of formative assessments and therefore there are many prospects for improvement.

Summative assessments in PGME (including General Surgery) are also seemingly done in a better manner in the United States (Dannefer, 2013; Epstein, 2007). The only difference in

the United States is that trainees take the American Board of Surgery In-Training Examination (ABSITE) instead of the CAGS in-training examination written in Canada ("General Surgery: Content Outline for the ABS In-Training Examination," 2013; Maciver, 2016; Ray et al., 2016). As well, more recently, the ABS has implemented the need to formally observe and assess six trainee operative procedures and six trainee clinical encounters as part of the summative ITER (Meyerson et al., 2014). In the United States if the licensing examination is taken it too consists of a multiple-choice and oral component ("ABS Booklet of Information Surgery," 2015). The lack of summative technical and non-technical performance assessments that occur outside of the ITER also currently plague their training system.

United Kingdom

In the UK, from an assessment standpoint, PGME is further ahead than their Canadian and American counterparts, particularly for formative assessments. As discussed in section 1.3.8 above, although the UK system has focused less on restructuring their education system towards CBME, their focus has been on overhauling their current assessment system, indirectly changing to a more CBME system ("General Surgery," 2015; "Good Medical Practice," 2014; "Good Surgical Practice," 2014). The UK system is structured around both formative and summative assessments (J. D. Beard, 2008; Shalhoub et al., 2014). The specifics of the formative assessments differ depending on the specialty, but the modalities used are the same (Rees et al., 2014; Shalhoub et al., 2014). Similarly, the summative assessments differ depending on the specialty, but culminate in the completion of a certification examination composed of written, oral and OSCE components ("Intercollegiate Specialty Examination in General Surgery," 2015; "Primary and Final FRCA Examination Regulations," 2015; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). The main concern seen by some in the UK is that although they have unveiled multiple strategies to be used for these assessments, evidence to support the interpretation of the results is lacking (Ali, 2013). Nevertheless, the transition of the UK is by far the most advanced in terms of assessment and their work to date forms the foundation of what is needed moving forward in Canada (and the United States) for successful competency-based assessment incorporation, discussed in section 1.5.1 below.

In the General Surgery specifically in the UK, formative assessments predominantly occur in a workplace setting (J. Beard, Rowley, Bussey, & Pitts, 2009; Eardley, Bussey, Woodthorpe, Munsch, & Beard, 2013; "General Surgery," 2015). The various modalities employed to assess trainees and provide feedback include: Case Based Discussions (CBD), Procedure Based Assessments (PBA), Direct Observations of Procedural Skills in surgery (DOPS), Clinical Evaluation Exercises (CEX) and Multi-Source Feedback (MSF) ("General Surgery," 2015). For all of these formative assessments, it is the responsibility of the trainee to initiate feedback and collect a predefined number of encounters for each type ("General Surgery," 2015). CBDs evaluate knowledge and judgment and are carried out between a trainee and staff surgeon where a case the trainee took part in is discussed, analyzed and avenues for improvement are sought ("General Surgery," 2015). PBAs focus on technical performance of index operative procedures and where along a continuum trainees lie in terms of being able to complete them independently ("General Surgery," 2015). PBAs are evaluated by staff surgeons and these PBAs provide an opportunity to assess the other competencies/domains within the UK CBME framework ("General Surgery," 2015). DOPS occur at earlier training levels (CT/ST 1 or 2) compared to PBAs and assess technical performance on more basic tasks or parts of larger procedures, again evaluated by staff surgeons ("General Surgery," 2015). CEXs evaluate all of the competencies/domains within the UK framework, by having a staff surgeon observe a trainee interacting with a patient during a clinical encounter ("General Surgery," 2015). MSF is a 360 degree evaluation of a trainee's professional/teamwork abilities where feedback is sought from various avenues including staff surgeons, nursing teams and allied health teams ("General Surgery," 2015). As can be seen by the various modalities utilized in the UK, formative assessments form a major part of the transition to CBME (Ali, 2013). The potential shortcomings in the UK, are the need for trainees to seek out staff surgeons for the various formative assessments – which creates an administration burden on their part and may cause the trainees to preferentially select more lenient assessors and/or a specific type of formative assessment (DOPS over CEX in more junior trainees and PBAs over CBDs in more senior trainees) (Phillips et al., 2015; Shalhoub, Santos, Bussey, Eardley, & Allum, 2015). Finally, as mentioned previously, there is also concern that more evidence is needed to properly interpret the results of these formative assessments (Ali, 2013). With further such substantiation coming out more recently in the work by Shalhoub *et al*., and Torsney *et al*. in their review (Shalhoub et al., 2014; Torsney, Cocker, & Slesser, 2015).

Summative assessments in the UK, more closely mirror those in Canada and the United States. The two major forms of summative assessments are the Annual Review of Competence Progression (ARCP) and the FRCS examination ("General Surgery," 2015; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). The ARCP is similar to the ITER evaluation completed in Canada and the United States, but it occurs at the end of each training year with the program director (Eardley et al., 2013). At the ARCP, evidence from the various formative assessments is collected and a decision about trainee progression or remediation is made (Eardley et al., 2013). The FRCS examination is the certification examination completed at the end of General Surgery training, which assesses knowledge and judgment predominantly ("Intercollegiate Specialty Examination in General Surgery," 2015). As in Canada and the United States the major assessment shortcomings in the UK include a lack of a summative evaluations specifically focusing on technical performance. As well as the absence of a summative assessment focusing on non-technical performance.

1.5 Assessment in the context of competency-based medical education

1.5.1 Summary of deficiencies in the current assessment systems

As can be seen above, in General Surgery all three jurisdictions for the most part focus their assessments on knowledge and judgment at both the formative, but particularly the summative time-points. The UK has incorporated many modalities into formative trainee feedback, which focus on all of the competencies/domains within their CBME framework; the major drawback being a lack of evidence to support their use (Ali, 2013; Shalhoub et al., 2014; Torsney et al., 2015). In Canada and the United States formative feedback remains relatively informal and unstructured, with some incorporation of the various competencies/domains within their specific CBME frameworks ("Common Program Requirements ", 2016; J. R. Frank, Snell, L.S., Sherbino, J, 2014). Summative assessments across all three jurisdictions include ITER/ARCP sessions as well as in-training and certification examinations. There is currently a paucity of summative assessments evaluating technical performance. As well, a void exists in evaluating non-technical performance and the modalities currently utilized may incorporate significant recall bias (Ginsburg et al., 2013).

In the section to follow, the focus will be on the important concepts that are needed to create competency-based assessments, which are now required given the recent transition to

CBME ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015; "Good Medical Practice," 2014; Holmboe et al., 2015; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014; "Tomorrow's Doctors," 2009). The emphasis will be on both the formative and summative side.

1.5.2 Milestones as central tenets in competency-based assessments

Central to competency-based assessments is the idea of milestones, which are specific attributes, expected of a trainee at a defined stage of training (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Iobst et al., 2010). These milestones are being created to direct learning through various types of assessments (J. R. Frank, Snell, L.S., Sherbino, J, 2014). They will help trainees focus their education and guide training programs in understanding whether trainees have attainted a particular ability (by reaching a specific milestone) (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Iobst et al., 2010).

In Canada these milestones are currently defined in a generic manner and apply to any medical or surgical specialty, with program specific milestones currently under development (J. R. Frank, Snell, L.S., Sherbino, J, 2014). In the United States these milestones have already been developed for each specialty (Green et al., 2009; Holmboe et al., 2015; Swing et al., 2013). In the UK, the term milestone is used less often, however the basis for progression in their training system is predicated on achieving specific competence levels at defined time-points, in keeping with the basis for the Canadian and American milestone projects ("General Surgery," 2015). Regardless of the jurisdiction, evaluations around these milestones will form the currency for feedback and progression within CBME from both a formative and summative standpoint (Holmboe et al., 2010). Although they have been or are in the process of being developed across all three jurisdictions, prior to their full implementation there are a few concepts that first need to be addressed.

1.5.3 Concepts relevant to both formative and summative competency-based assessments

Regardless of the modalities utilized or the purpose of the assessment, there must be evidence to support the interpretation of the assessment results (Korndorffer, Kasten, & Downing, 2010). In a traditional sense this was referred to as validity and reliability – with

multiple forms falling under each of the categories (i.e. content/face validity and inter-rater/intra-rater reliability for example) (D. A. Cook & Beckman, 2006; Jelovsek, Kow, & Diwadkar, 2013). In a more modern sense, first introduced by Messick, there is now a single unitary concept of validity, referred to solely as construct validity (D. A. Cook & Beckman, 2006; D. A. Cook & Lineberry, 2016; S. Messick, 1989; Messick, 1995). Reliability no longer stands alone but rather falls under this unitary validity concept, and instruments or modalities are not said to be valid themselves, instead evidence is collected to support their use for a specific purpose (D. A. Cook & Beckman, 2006; D. A. Cook & Lineberry, 2016; S. Messick, 1989; Messick, 1995). In this new conceptual framework of validity, there are five main sources of evidence that may be collected to support the interpretation of the assessment results and they include: content, response process, internal structure, relationship to external variables and consequences (D. A. Cook & Beckman, 2006; D. A. Cook & Lineberry, 2016; D. A. Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014).

Content evidence is comprised of multiple components and in essence refers to how the makeup of the assessment instrument relates to the construct being measured (i.e. technical performance, knowledge, communication, professionalism etc.) (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995). In particular, the creators and users of these assessment instruments need to be appropriately qualified, the assessment items must themselves be representative of the domains for which they were created and there must be a logical relationship between content tested and the achievement domain (S. M. Downing, 2003; Ghaderi et al., 2015). Response process refers to analyzing the causes of variance not relevant to the construct being measured and minimizing them, with these most notably including variance related to assessment administration and the methods by which data are collected and stored (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995) Internal structure refers to how the items of an instrument fit the underlying construct; in essence evaluating the reproducibility (reliability) of the results (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995) Relationship to external variables refers to the association that exists between scores achieved using the assessment instrument and external variables that measure the same construct (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995). For example, trainees that perform well in teaching rounds answering questions, should also perform well on the in-training examination evaluating the

same topics. Finally, consequences refers to the impact the results will have on trainees and how they will be used – whether the results for example, will be utilized for trainee feedback or to make decisions about trainee progression (D. A. Cook & Beckman, 2006; D. A. Cook & Lineberry, 2016; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995).

1.5.4 Concepts specific to formative competency-based assessments

As discussed in section 1.4.3, the UK has implemented many formative competency-based assessments into surgical training. These assessments contain some, but not all of the important concepts central to the implementation of such formative evaluations.

The major facets imperative to formative assessments within CBME includes: 1) assessor (staff surgeon) training, 2) assessment variability, and 3) program level initiation.

The staff surgeons completing formative evaluations must be trained in doing so (J. Beard et al., 2009). This training should be centered around the role such assessments play in trainee education, how to carry out the assessments (assign scores) and how to provide effective feedback (J. Beard et al., 2009; M. Feldman, Lazzara, Vanderbilt, & DiazGranados, 2012; J. Norcini et al., 2011). Specifically, staff surgeons must understand the scope of each assessment instrument, its purpose (which competencies it does/does not assess) and the appropriate location for its use (in the OR, outpatient clinic, teaching rounds etc.) (D. A. Cook et al., 2014; M. Feldman et al., 2012; "General Surgery," 2015). Furthermore, these surgeons must be able to actually use the assessment instrument – having a firm grasp on the specific constructs under evaluation within each scale item and how to minimize biases or other factors, which may affect the accuracy of their results (M. Feldman et al., 2012; Weitz et al., 2014). Finally, as these assessments are formative, the surgeons must provide effective feedback necessary for trainee improvement (Ramani & Krackov, 2012). Aspects of such feedback include ensuring it is timely, having it occur directly after an observed encounter, asking the trainee for their self assessment prior to proceeding, verbalizing specific areas for improvement and finally confirming the trainee understands the feedback and can address it moving forward (Gonzalo et al., 2014; Ramani & Krackov, 2012).

Assessment variability is comprised of assessing a variety of competencies, utilizing various assessment strategies, using multiple assessors and evaluating trainees on multiple

occasions (C. Carraccio et al., 2015; Holmboe et al., 2010). The various assessments should evaluate each competency/domain within the specific jurisdiction's CBME. There may be a tendency at the assessor and trainee level to preferentially continue to evaluate and be evaluated on competencies/domains, which are easier to measure such as knowledge and judgment. This is an area where the UK has found some difficulty with trainees favoring assessments geared at specific competencies/domains over others (Shalhoub et al., 2015). Moreover, not only should the various competencies/domains play a major role in assessment, so too should various modalities (as in the UK using PBAs, CBDs, DOPS, CEX) (Holmboe et al., 2010; Swing, Clyman, Holmboe, & Williams, 2009). Finally, trainees should be assessed on multiple occasions, by multiple assessors to ensure that their observed performance of a particular competency/domain is a true reflection of their ability (J. R. Wilkinson et al., 2008). This concept was nicely illustrated by Marriott *et al*. and Crossley *et al*. evaluating the technical and non-technical skills of surgical trainees respectively, depicting that at least three assessors and three procedures per trainee were required to produce reliable results (Crossley, Marriott, Purdie, & Beard, 2011; Marriott, Purdie, Crossley, & Beard, 2011). Guldbrand Nielsen *et al*. recently substantiated these findings for cardiac trainees undertaking thoracic imaging procedures (Guldbrand Nielsen, Jensen, & O'Neill, 2015).

Program level initiation or at least program level support is imperative for the implementation of formative assessments (J. Norcini et al., 2011). In the UK currently, competency-based assessments are for the most part initiated by the trainees ("General Surgery," 2015; Shalhoub et al., 2014). This has shown to burden the trainees and presumably negatively impact their education and outlook towards such assessments (Bindal, Wall, & Goodyear, 2011; Pereira & Dean, 2009, 2013). This has subsequently resulted in the trainees trying to circumvent the real purpose of formative assessments by focusing solely on collecting the appropriate number required by their training program in a manner that is not conducive to learning (batching assessments at the end of a rotation to complete them, but when no further improvement with that trainer may occur) (Ali, 2013; Bindal et al., 2011). It should be the focus and responsibility of the training program to ease the burden of their trainees by off loading some of the other non-educational components, which may be impacting their ability to focus on and learn from such formative assessments. Alternatively, the training program may shift some of the assessment initiation on to the part of the staff surgeons.

1.5.5 Concepts specific to summative competency-based assessments

Current summative assessments predominantly focus on knowledge and judgment across the three jurisdictions as discussed above in section 1.4.2 – 1.4.3. None of these jurisdictions have focused their attention to technical and non-technical performance assessments in the summative context (that do not occur as part of the ITER evaluation). Prior to the implementation of such summative assessments however, criterion-referenced performance standards that can credibly and reliably separate competent from non-competent trainees, must be created (Cizek, 1993; Cizek & Bunch, 2007f; J. J. Norcini, 2003; J. J. Norcini, Shea, JA, 1997).

The specifics of standard setting are discussed thoroughly in sections 1.9.1 – 1.9.2. Suffice it to say that although standards are imperative for the application of summative assessments into surgical training and this knowledge being echoed in the literature, their creation and implementation continues to be scarce as outlined in sections 1.9.6 – 1.9.8 (Holmboe et al., 2010; Kogan & Holmboe, 2013). Like their formative assessment counterparts above, these summative assessments must also utilize trained assessors and demonstrate assessment variability (C. Carraccio et al., 2015; Crossley et al., 2011; Holmboe et al., 2010; Marriott et al., 2011; Swing et al., 2009; J. R. Wilkinson et al., 2008).

1.6 Assessment of technical performance

1.6.1 Impetus for assessing technical performance

As described in the above sections, the traditional training and assessment frameworks focus primarily on knowledge and judgment – with a void in assessing technical performance. With the implementation of CBME and its focus on workplace assessments, there is a real need to delineate the methods by which technical competence (irrespective of whether it is done for formative or summative reasons) is currently assessed in surgical trainees. Furthermore despite the transition to CBME, a clear definition of competence is lacking, with various terms including skill, proficiency and competency often used interchangeably. Finally, determining whether performance standards have been utilized to differentiate competent from non-competent trainees in surgery is also imperative. These three voids go on to form Aim # 1 in this thesis.

1.6.2 Instruments for technical performance assessment

Introduction

There has been a push for the evolution of surgical education programs in the last decade, with respect to how surgical residents are trained, at the local, national and international levels (Debas et al., 2005; Pellegrini, 2006; Peracchia, 2001; Warnock, 2012). This has come from various stakeholders; including the training programs themselves, hospital systems, licensing authorities, governments, and the public, in response to recent changes (Darzi, 2008; Darzi, Datta, & Mackay, 2001; Kohn, 1999; Pellegrini, 2012; Pellegrini, Warshaw, & Debas, 2004; Sachdeva et al., 2007; R. Smith, 1998). These changes include work hour restrictions in various jurisdictions (Barden, Specht, McCarter, Daly, & Fahey, 2002; Drolet, Sangisetty, Tracy, & Cioffi, 2013; Villaneuva, 2010), fiscal constraints that limit operating room accessibility (Bridges & Diamond, 1999; Good, Khan, Kiely, & Brady, 2013; Hosler et al., 2012), and increased litigation against physicians (Berry, 2006), all of which decrease operative and educational opportunities (Ziv, Wolpe, Small, & Glick, 2003).

In response to these changes and as a means of maintaining their high standards of education, various governing bodies including the ACGME ("Next accreditation system (NAS) milestones," 2012), the RCPSC ("CanMEDS," 2013), and many others worldwide ("Intercollegiate surgical curriculum programme," 2013; "Surgical Education and Training Policies," 2013), have mandated competency-based education paradigms, with specific attributes that must be completed during training. The new CanMEDS 2015 framework, which is currently in development, will continue to feature CBME with a new priority of milestones, embedded within each of the seven attributes and the overall goal to “implement outcomes driven education and assessment to ensure physicians possess the ability they need for every stage of their career” ("Competency-based medical education," 2013).

The assessment of these milestones however, with respect to technical competence within CBME has proven difficult. There continues to be a poor working definition of technical competence and a unifying definition among researchers and educators remains elusive. The literature is also fraught with the terms competence and proficiency being used interchangeably. Finally, there is very limited literature describing the current methods used to assess technical competence in residents.

With the ongoing application of competency-based education initiatives in residency programs ("Competency-based medical education," 2013), and the shift in medical education overall ("CanMEDS," 2013; "Intercollegiate surgical curriculum programme," 2013; "Next accreditation system (NAS) milestones," 2012; "Surgical Education and Training Policies," 2013), the need to define technical competence and develop competency-based assessments to complement and evaluate these educational initiatives is imperative.

The objective of this review is to systematically examine the methods by which technical competence is assessed in surgical trainees and to evaluate the validity and reliability of these methods and document the standard setting strategies (J. J. Norcini, 2003; Schindler, Corcoran, & DaRosa, 2007; T. J. Wilkinson, Newble, & Frampton, 2001) utilized. A secondary objective includes assessing the quality of evidence to determine the transferability across other institutions and our confidence in the estimate of effect. Lastly, an evaluation of the literature is conducted with regards to the overarching definition of technical competence and how the published literature differentiates between competence and proficiency.

Methods

Protocol

A systematic review protocol was created in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement (Moher, Liberati, Tetzlaff, Altman, & Group, 2009).

Eligibility

Study characteristics as per the PICOS (population, intervention, comparison, outcome, setting) model, included residents or interns, in any surgical postgraduate program (including Obstetrics and Gynecology), whose technical competence (as defined by each individual study) was assessed. Studies that were eligible, irrespective of publication status, included RCTs and observational studies, such as: non-randomized trials (quasi-experimental), cohort studies, case-control studies, case-series and cross-sectional studies.

Information sources

A search was performed in Ovid MEDLINE (R) from 1946 to July week 3, 2013, Embase Classic + Embase 1947 to week 31, 2013, PsychINFO 2002 to August week 1, 2013 and the Cochrane Library 1993 to August 6, 2013.

Search

One author (P.S.) and a St. Michael's Hospital librarian, Toronto, Canada, searched the above databases independently. The medical subject headers (MESH) used included: 'surgical procedures, operative' 'specialties, surgical', 'laparoscopy', 'internship and residency', 'task performance and analysis' and 'educational measurement'. Key terms used included: 'surgery', 'neurosurgery', 'operative', 'laparoscope', 'minimally invasive', 'resident', 'intern', 'competency-based education', 'measure', 'evaluate', 'assess', 'criteria', 'benchmark', 'gauge', 'clinical competence', 'competent', 'capability', 'expertise', 'skills', 'motor skills', 'technique' and 'proficiency'. The researchers also used appropriate variations to account for word variations and plurals.

Titles and abstracts were initially reviewed and relevant publications were retrieved and accessed for full-text review. Reference lists of relevant publications were also hand searched to ensure studies were not omitted by the computer search strategy, until no additional relevant publications were discovered.

Study selection/inclusion criteria/exclusion criteria

For inclusion in the systematic review, publications needed to assess the technical competence of surgical trainees in any surgical subspecialty, including Obstetrics and Gynecology. Only original articles were included. Opinion letters, case reports, reviews and letters to the editor were excluded. Studies were limited to English language.

Two authors (P.S. and M.L.) reviewed articles independently and any disagreements were resolved through discussion and consensus.

Data extraction/synthesis of results

Data extraction was completed in a systematic fashion. The included studies were evaluated for number of participants, types of participants (either categorical or preliminary surgery residents), the methods by which they assessed trainee technical competence, and the primary outcome(s) in each study.

Examination of validity and reliability of assessment methods and standard setting approaches

The publications were assessed for assessment method validity, reliability and where applicable, whether standard setting (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001) approaches were utilized. This was accomplished by gathering information directly from the included studies themselves or referenced papers in the included studies, if they used previously developed methods in their unchanged form.

Quality assessment

The publications were also assessed by two authors (P.S. and M.L.) for quality using the GRADE system (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008). The GRADE (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008) classification was utilized to assess the transferability of the assessment methods and to document how well the study was carried out, the results reported and the likelihood that further research would change our confidence in the estimate of effect.

Results

Included studies

The initial search strategy identified 6814 studies (Appendix 1). After screening the titles and abstracts based on our inclusion and exclusion criteria, 291 studies were selected. An additional 20 studies were added after relevant bibliographies were hand searched, bringing the total number of studies that underwent full-text review to 311. Of these 311 studies, 85 were ultimately included in this systematic review involving 2369 surgical residents (Figure 2). Studies were excluded for the following reasons; Review/editorial/opinion/dialogue publications (27), wrong focus (i.e. professional competence) (107), wrong study group (i.e. medical students)

(32), assessment of technical skill, not technical competence (32), measurement tool/scale not provided (15), duplicates (5) and those with no full text publications, only abstracts with insufficient information (8). Of the 85 included studies, the method by which they assessed trainee technical competence can be broken down into five main groups based on our findings; Likert scales (37), benchmarks (31), binary outcomes (11), novel tools (4) and surrogate outcomes (2). These groups are summarized below and depicted in Table 1. Also summarized below are the platforms; live versus video recorded the included studies utilized to complete their assessments.

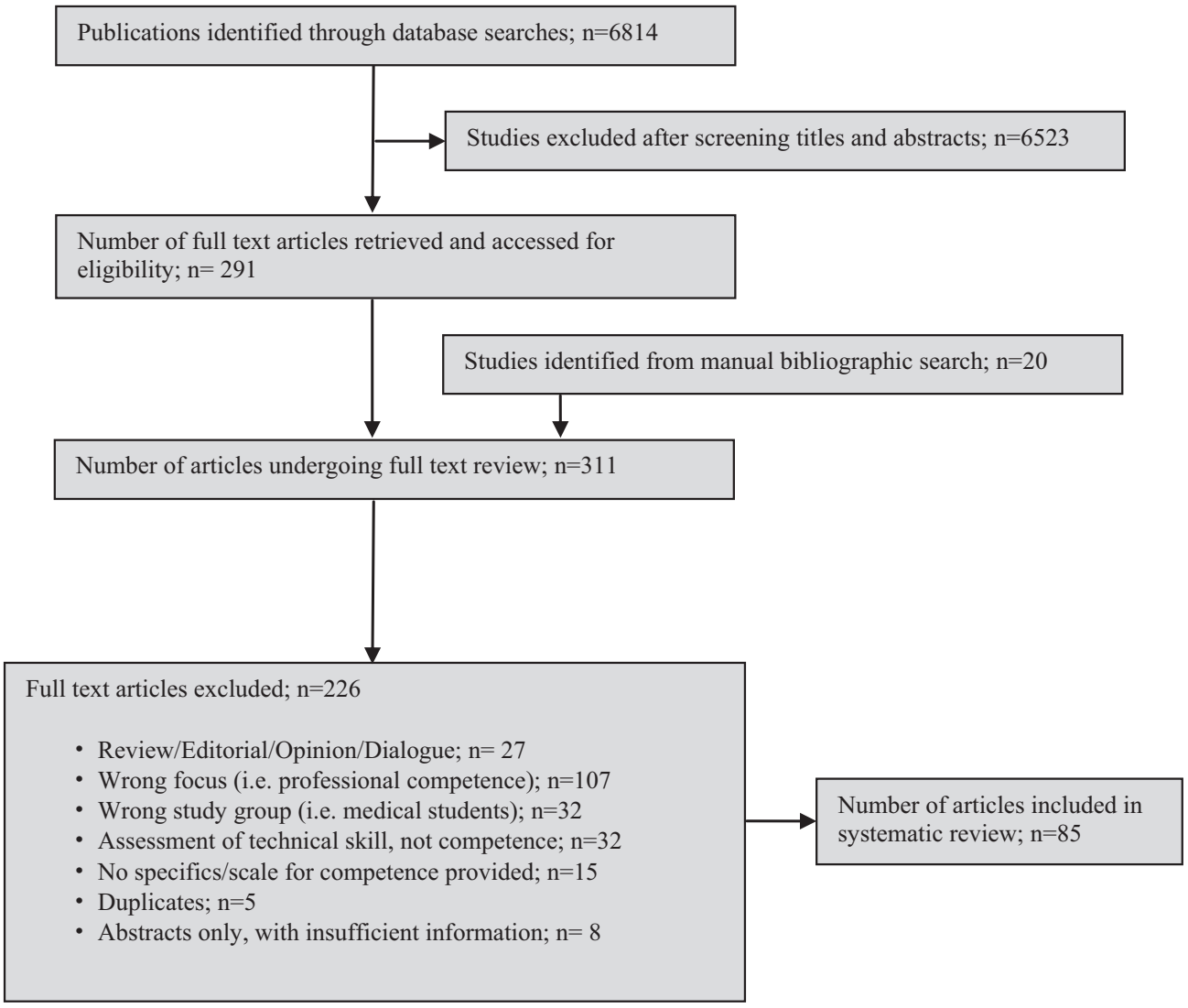

Figure 2. Flow diagram depicting systematic review strategy.

Table 1. The methods by which technical competence is assessed in surgical trainees, categorized by method of assessment

Assessment method	Number of studies	Number of participants
1) Likert scales	**37**	**819**
Global rating scale	7	108
Task-specific scale	3	74
Combined*	27	637
2) Benchmarks	**31**	**1113**
Expert benchmarks	28	790
Relative benchmarks	3	323
3) Binary outcomes	**11**	**279**
Overall binary outcomes (Y/N)	7	214
Task-specific binary outcomes (Y/N)	4	65
4) Novel tools	**4**	**147**
Novel tools	3	134
Error analysis	1	13
5) Surrogate Outcomes	**2**	**11**
Cecal intubation	1	10
Tensile strength	1	1
TOTAL	**85**	**2369**

*Studies which used a combined approach incorporating a global rating scale and task-specific scale in their assessment

Validity and reliability of assessment methods and standard setting approaches

We evaluated each included study for validity (face, content, construct, external, predictive), reliability (inter-rater, intra-rater, internal consistency), and standard setting (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001) approaches (Angoff, contrasting groups, Hofstee, receiver operating characteristic (ROC) curve). The methods were either validated and their reliability documented in the paper included in this systematic review, or as was the majority, in previous papers published using the assessment method in a different setting. Results from previous studies were only included if they used the method in its original form, without any change from the time it was validated and its reliability assessed (Table 2). In a few studies the type of validity and reliability measures documented were not further defined (i.e. construct versus face validity) and these are presented in our paper as validity type not described ($V^{Ø}$) or reliability type not described ($R^{Ø}$) respectively. Similarly the included studies were assessed for standard setting approaches (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001). In this regard, two studies described a previously utilized standard setting approach (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001), the ROC to determine a competency "cut-off" score that optimizes sensitivity and specificity for differentiating competent versus non-competent trainees (Fraser et al., 2003; Swanstrom, Fried, Hoffman, & Soper, 2006), or predicting performance at or above the level of an experienced surgeon (Sroka et al., 2010) (McCluney et al., 2007). Another study established minimum "cut-off" scores to assess competency in Gynecologic trainees using the modified Angoff approach as the primary standard setting measure and the contrasting groups method and Hofstee method for confirmation (Jelovsek et al., 2010). Finally, Beard *et al*. (J. D. Beard, Education, Training Committee of the Vascular Society of Great, & Ireland, 2005) used the contrasting groups method to differentiate competent versus non-competent Vascular trainees.

Table 2. Psychometric properties (validity, reliability) * of assessment methods and standard setting approaches

Assessment method	Psychometric properties	Standard Setting
1) Likert scales		
Global rating scale	IRR, IAR, IC, CV, CCV, COV	ROC
Task-specific scale	IRR, IC, CV,	Not documented
Combined	IRR, IAR, IC, CV, CCV, COV, FV, $V^{Ø}$	Modified Angoff‡, CGM
2) Benchmarks		
Expert benchmarks	IRR, R^{T}, IC, CV, CCV, COV, FV, PV, EV, $V^{Ø}$	ROC
Relative benchmarks	IRR, $R^{Ø}$, CV, CCV	Not documented
3) Binary outcomes		
Overall binary outcomes (Y/N)	IRR, IC, CV, FV	Not documented
Task-specific binary outcomes (Y/N)	R^{C}, CV, V^{C}, $V^{Ø}$	Not documented
4) Novel tools		
Novel tools	IRR, R^{G}, IC, CV	Not documented
Error analysis	IRR, CV	Not documented
5) Surrogate Outcomes		
Cecal intubation	Not documented	Not documented
Tensile strength	Not documented	Not documented

*Psychometric properties of assessment methods include: IRR: Inter-rater reliability, IAR: Intra-rater reliability, IC: Internal consistency, R^{G}: Reliability via genaralizability analysis, R^{C}: Reliability coefficient, R^{T}: Test –retest reliability, $R^{Ø}$: Reliability type not described, CV: Construct validity , VC: Validity via cognitive task analysis, FV: Face validity, CCV: Concurrent validity, COV: Content validity, EV: External validity, PV: Predictive validity $V^{Ø}$: Validity type not described ROC: Receiver operator curve, CGM: Contrasting groups methodology ‡ This study performed both a modified Angoff methodology as the primary standard setting approach and the contrasting groups and Hofstee methodology for confirmation

Methodological quality

The quality of the included studies as assessed by the GRADE classification (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008) is summarized in Table 3. Of the 85 included studies, 15 were RCTs and their overall quality for the primary outcome examined in each study was found to be high, whereas 70 were observational studies and their overall quality for the primary outcome examined in each study was found to be low.

Table 3. GRADE* classification of included studies, organized by assessment methodology and study type

Number of studies (Participants)	Limitations	Precision	Consistency	Directness	Publication bias	GRADE**
Likert scales						
9 RCT (n= 181)	Serious	Imprecision	Consistent	Direct	Not detected	High (++++)
28 Obs (n= 638)	Serious	Imprecision	Inconsistent	Direct	Not detected	Very low (+)
Benchmarks						
3 RCT (n= 85)	Not Serious	No imprecision	Consistent	Direct	Not detected	High (++++)
28 Obs (n= 1028)	Serious	Imprecision	Consistent	Direct	Not detected	Low (++)
Binary outcomes						
2 RCT (n= 54)	Serious	No imprecision	Consistent	Direct	Not detected	Undetermined‡
9 Obs (n= 225)	Serious	No imprecision	Consistent	Direct	Not detected	Low (++)
Novel tools						
1 RCT (n=13)	Not serious	No imprecision	Consistent	Direct	Not detected	High (++++)
3 Obs (n=134)	Serious	No imprecision	Consistent	Direct	Not detected	Very low (+)
Surrogate Outcomes						
0 RCT						
2 Obs (n= 11)	Serious	Imprecision	Consistent	Direct	Not detected	Very low (+)

* GRADE (Grading of Recommendations Assessment, Development and Evaluation) of evidence for the primary outcome identified in each study
** Overall GRADE: High quality ++++, Moderate quality +++, Low quality ++, Very low quality +
‡ Of the two RCTs included, the quality was undetermined, one study was High quality (++++), and the other Moderate quality (+++)
RCT: Randomized controlled trial
Obs: Observational study (cross-sectional, cohort, case-series, pre-post quasi-experimental

Likert scale assessments

These assessments all employed the original scale used in the Objective Structured Assessment of Technical Skill (OSATS) examination, developed by Reznick *et al*. (Reznick, Regehr, MacRae, Martin, & McCulloch, 1997), or a modification of that scale, which was found to best fit the assessment goals in a particular specialty in that particular publication. Of the 37 studies in this category, seven utilized a global rating scale, three utilized a task-specific scale and 27 use a combined approach.

The global rating scale in the included studies assessed technical competence across multiple disciplines including, General Surgery (Hernandez et al., 2004; Hogle, Widmann, Ude, Hardy, & Fowler, 2008; Palter, Grantcharov, Harvey, & Macrae, 2011; Sroka et al., 2010), Gynecology (Chou, Bowen, & Handa, 2008) Otolaryngology (Syme-Grant, White, & McAleer, 2008), and Cardiac Surgery (Hance et al., 2005). They used various platforms to acquire technical competence information and then utilized the global rating scale for assessment. These platforms included recorded (Hogle et al., 2008; Palter, Grantcharov, et al., 2011; Sroka et al., 2010) or live (Chou et al., 2008; Hernandez et al., 2004; Syme-Grant et al., 2008) assessments that were performed intra-operatively via a laparoscopic approach, open approach or a da Vinci robot or live assessments that were performed on standardized bench model (Hance et al., 2005).

The task-specific scale in this section is in fact a likert scale with task-specific components. It is not to be confused with a task-specific checklist found in the section 'Binary outcome assessments'. The task-specific scale in the included studies evaluated technical competence in Urology (Maizels et al., 2008), Otolaryngology (Francis, Masood, Laeeq, & Bhatti, 2010) and one of the studies evaluated surgery residents in the Emergency Room setting (Fargo, Edwards, Roth, & Short, 2011) (residents were from Orthopedics, Obstetrics, General Surgery, Otolarygology and Emergency Medicine, classified in this study as surgical based on their hands-on exposures). The platforms utilized were live assessments performed intra-operatively or intra-procedurally via an open approach (Fargo et al., 2011; Francis et al., 2010; Maizels et al., 2008).

The combined use of both scales was most in keeping with the original OSATS (Reznick et al., 1997) examination and employed both a global rating scale and task-specific rating scale or checklist (the task-specific checklist was included in this section as well, instead of the section

'Binary outcome assessments' to keep the original OSATS (Reznick et al., 1997) format/exam together and for ease of explanation). The combined approach included studies in General Surgery (Orzech, Palter, Reznick, Aggarwal, & Grantcharov, 2012; Palter & Grantcharov, 2012; Palter, Orzech, Reznick, & Grantcharov, 2013; Vassiliou et al., 2005), Gynecology (Jelovsek et al., 2010), Ophthalmology (Taylor, Binenbaum, Tapino, & Volpe, 2007), Orthopedics (Howells, Gill, Carr, Price, & Rees, 2008; Insel, Carofino, Leger, Arciero, & Mazzocca, 2009), Vascular Surgery (J. D. Beard, Choksy, & Khan, 2007; J. D. Beard et al., 2005; Pandey et al., 2006), and Otolaryngology (A. Ahmed, Ishman, Laeeq, & Bhatti, 2013; Fleming, Kapoor, Sevdalis, & Harries, 2012; Ishman et al., 2010; Laeeq et al., 2009; Laeeq et al., 2010; Lin et al., 2009; Malik, 2011; Stack, Siegel, Bodenner, & Carr, 2010; Varela DA, 2011). A few of the studies (Chipman & Schmitz, 2009; Khan, Bann, Darzi, & Butler, 2003; Rooney, Hungness, Darosa, & Pugh, 2012; Sanfey et al., 2010; Wade & Webb, 2013) evaluated the technical competence of categorical or preliminary residents from various specialties performing basic tasks (i.e. skin suturing) or generic procedures (i.e. chest tube insertion). Two studies evaluated the technical competence of categorical or preliminary residents from various specialties performing complex tasks (Gauger et al., 2010; Mitchell et al., 2011) (laparoscopic cholecystectomy, end-to-end vascular anastamosis). The platforms utilized included recorded assessments of procedures performed intra-operatively via a laparoscopic (Orzech et al., 2012; Palter & Grantcharov, 2012; Palter et al., 2013), open (J. D. Beard et al., 2005; Jelovsek et al., 2010; Mitchell et al., 2011; Pandey et al., 2006; Taylor et al., 2007), or combined laparoscopic/open approach (Sanfey et al., 2010). Live assessments of procedures performed intra-operatively or intra-procedurally via an open (J. D. Beard et al., 2007; Chipman & Schmitz, 2009; Howells et al., 2008; Stack et al., 2010) or laparoscopic/endoscopic approach (A. Ahmed et al., 2013; Gauger et al., 2010; Ishman et al., 2010; Laeeq et al., 2010) and a laparoscopic procedure were assessed via both a live and recorded manner (Vassiliou et al., 2005). The platforms also utilized recorded assessments on procedures performed on standardized bench models via an open (Khan et al., 2003; Wade & Webb, 2013) approach or live assessments performed on standardized bench models via an open or endoscopic approach (Fleming et al., 2012; Insel et al., 2009; Laeeq et al., 2009; Lin et al., 2009; Rooney et al., 2012; Varela DA, 2011). Finally, one study assessed live performance on a virtual reality simulator (Malik, 2011).

Benchmark assessments

In this review, 31 studies utilized benchmark measures to assess technical competence; 28 of these were compared to expert benchmarks and three were compared to relative (resident referenced) benchmarks.

The expert benchmark studies assessed competence in multiple disciplines including General Surgery (Brunner et al., 2005; Grantcharov & Funch-Jensen, 2009; Jenison, Gil, Lendvay, & Guy, 2012; Korndorffer, Dunne, et al., 2005; Lendvay et al., 2013; Mashaud et al., 2010; Salgado, Grantcharov, Papasavas, Gagne, & Caushaj, 2009; Selvan, 2011; Stefanidis, Acker, & Greene, 2010; Sutton et al., 2013; Swanstrom et al., 2006; van Dongen et al., 2011; Verdaasdonk, Dankelman, Lange, & Stassen, 2008; Zendejas, Cook, Hernandez-Irizarry, Huebner, & Farley, 2012), Gynecology (Jenison et al., 2012; Kolkman, Van de Put, Van den Hout, Trimbos, & Jansen, 2007; Lendvay et al., 2013; A. K. Moore, Grow, Bush, & Seymour, 2008), Otolaryngology (Fried et al., 2012) and Urology (Jenison et al., 2012; Lendvay et al., 2013). The majority of studies assessed the technical competence of categorical and/or preliminary surgical trainees performing a basic task (i.e. knot tying) or generic procedure (i.e. laparoscopic suturing) without a clear discipline focus or breakdown of participating residents (Brinkman, Buzink, Alevizos, de Hingh, & Jakimowicz, 2012; Conway, 2011; Debes, Aggarwal, Balasundaram, & Jacobsen, 2012; Goova et al., 2008; Korndorffer, Scott, et al., 2005; Nugent et al., 2013; Perrenot, 2011; D. J. Scott, Goova, & Tesfay, 2007; Shane, Pettitt, Morgenthal, & Smith, 2008; Stefanidis et al., 2005; Stefanidis et al., 2006). The platforms utilized were live performances on virtual reality (Brinkman et al., 2012; Brunner et al., 2005; Conway, 2011; Grantcharov & Funch-Jensen, 2009; Lendvay et al., 2013; A. K. Moore et al., 2008; Nugent et al., 2013; Perrenot, 2011; Salgado et al., 2009; Shane et al., 2008; Sutton et al., 2013; van Dongen et al., 2011; Verdaasdonk et al., 2008), synthetic models (box trainers) (Debes et al., 2012; Jenison et al., 2012; Kolkman et al., 2007; Korndorffer, Dunne, et al., 2005; Korndorffer, Scott, et al., 2005; Mashaud et al., 2010; Stefanidis et al., 2010; Stefanidis et al., 2006; Swanstrom et al., 2006; Zendejas et al., 2012), simulators (Fried et al., 2012; Selvan, 2011), or a combination of synthetic models and a virtual reality platform (Stefanidis et al., 2005). Two studies utilized live assessments performed on standardized bench models (Goova et al., 2008; D. J. Scott et al., 2007).

Von Websky *et al*. collected an extensive database of trainees performing basic and procedural tasks mostly general surgery related, on a virtual reality platform and used the database to set competency benchmarks (von Websky et al., 2012). Brydges *et al*. assessed general surgery trainees performing various tasks using the Imperial College Surgical Assessment Device (ICSAD); a motion analysis system that tracks the three dimensional coordinates of a participant's hands to assess performance (Brydges, Classen, Larmer, Xeroulis, & Dubrowski, 2006). The scores of junior trainees were compared to senior trainees performing the same tasks to assess competency. Finally, Ganju *et al*. assessed neurosurgery residents performing tasks part of a validated course using a virtual reality platform, in the pre-call and post-call states (Ganju et al., 2012). The results from the pre-call state were established (set as a sort of benchmark), and the residents attempted to reach these benchmarks in the post-call state and any differences were noted (Ganju et al., 2012).

Binary outcome assessments

Binary assessments can take on a variety of forms. The studies included in this review evaluate either binary global competence outcomes (ability to perform a procedure overall) or binary task-specific competence outcomes (satisfying each task that together makeup a procedure). Of the 11 studies in this category, seven assessed binary global competence outcomes and four assessed binary task-specific outcomes.

The binary global competence outcomes evaluated procedures in General Surgery (Adrales, Chu, et al., 2003; Adrales et al., 2004; Adrales, Park, et al., 2003), Obstetrics and Gynecology (Goff et al., 2005), Otolaryngology (Bath & Wilson, 2007), and Trauma/Critical Care (M. Martin et al., 1998; Parent et al., 2010) where the trainees in this last group were either categorical /preliminary surgical interns, not presented by specialty. They utilized recorded (Adrales, Chu, et al., 2003; Adrales et al., 2004; Adrales, Park, et al., 2003) or live (Bath & Wilson, 2007; Goff et al., 2005; M. Martin et al., 1998; Parent et al., 2010) assessments performed intra-operatively via a laparoscopic or open approach.

The binary task-specific outcomes evaluated procedures in Ophthalmology (Taravella, Davidson, Erlanger, Guiton, & Gregory, 2011), and General Surgery (Proctor et al., 1998; Wilasrusmee, Lertsithichai, & Kittur, 2007). There was also a study which evaluated trainees

that were either categorical or preliminary surgical interns currently rotating through a Critical Care setting (Velmahos et al., 2004). They utilized live assessments performed intra-operatively or intra-procedurally via an open approach (Proctor et al., 1998; Taravella et al., 2011; Velmahos et al., 2004; Wilasrusmee et al., 2007).

Novel tools assessments

The majority of technical competence assessments in this review utilize methods that were originally designed to assess technical skill. There are, however, four novel tools included that have been developed more recently by institutions, whereby three directly assess technical competence and one employs error analysis to infer technical competence.

Miskovic *et al*. developed a generic technical competence tool for different laparoscopic colorectal surgery procedures. The scores were defined by the degree of intra-operative support the trainees required ranging from full support; 1 – Step done by trainer, to no support; 5 – Safe with no guidance and 6 – Could not be better (Miskovic, Wyles, Carter, Coleman, & Hanna, 2011). Miskovic *et al*. utilized live assessments performed intra-operatively via a laparoscopic approach (Miskovic et al., 2011). Similarly, Gofton *et al*. assessed a trainee's technical competence to independently perform a surgical procedure regardless of type and postgraduate year (Gofton, Dudek, Wood, Balaa, & Hamstra, 2012). Their tool assessed procedures in General Surgery and Orthopedics and scores were defined by the extent to which the staff surgeon was comfortable in allowing the trainee to be the primary operator, and how comfortable the staff was in leaving the operating room (Gofton et al., 2012). The scores ranged from 1 – staff having to do the procedure to 5 – staff presence was not required. Gofton *et al*. utilized live assessments performed intra-operatively via an open or laparoscopic approach (Gofton et al., 2012). In a study evaluating how many supervised examinations are required to achieve competence by surgical residents in fiberoptic sigmoidoscopy, Hawes *et al*. employed a gestalt or overall competence score that divided the residents into being able to perform the procedure independently or not (Hawes et al., 1986). The residents evaluated were categorical or preliminary surgical trainees and the platform utilized live assessments performed intra-procedurally (Hawes et al., 1986).

Ahlberg *et al*. used an error rating tool, divided into exposure errors, dissection errors and clipping and tissue division errors, while evaluated general surgery residents perform a laparoscopic cholecystectomy to infer differences in technical competence/performance (Ahlberg et al., 2007). They utilized live assessments performed intra-operatively (Ahlberg et al., 2007).

Surrogate outcome assessments

Two studies in this review assessed trainees by using surrogate markers to evaluate technical competence. Matsuda *et al*. evaluated how quickly surgical residents can learn to perform a colonoscopy by using cecal intubation rate and time as objective criteria to infer technical competence (Matsuda, 2012). The modality utilized by Matsuda *et al*. was intra-operative live assessment. Zetlitz *et al*. evaluated the mechanical characteristics of fresh flexor digitorum profundus tendon repairs that were preconditioned and distracted to failure. Biomechanical parameters including ultimate tensile strength, yield strength, 3mm gap force and stiffness were used as criteria to infer technical competence (Zetlitz, Wearing, Nicol, & Hart, 2012). The modality utilized in this study was a live assessment performed on a standardized bench model.

Definition of technical competence

Within the reviewed literature, most definitions of technical competence have embedded within them a minimum standard to safely perform a procedure or task independently (Adrales et al., 2004; A. Ahmed et al., 2013; Bath & Wilson, 2007; Fleming et al., 2012; Francis et al., 2010; Fried et al., 2012; Goff et al., 2005; Gofton et al., 2012; Hawes et al., 1986; Howells et al., 2008; Ishman et al., 2010; Jelovsek et al., 2010; Laeeq et al., 2009; Laeeq et al., 2010; Lin et al., 2009; M. Martin et al., 1998; Miskovic et al., 2011; Pandey et al., 2006; Proctor et al., 1998; Stack et al., 2010; Swanstrom et al., 2006; Taravella et al., 2011; Taylor et al., 2007; Vassiliou et al., 2005).

Differentiating competence from proficiency in the reviewed literature, demonstrated that a proficient performer is one who approaches expertise and is thus at an advanced stage compared to a competent performer (Ahlberg et al., 2007; A. Ahmed et al., 2013; Brydges et al., 2006; Debes et al., 2012; Fried et al., 2012; Goova et al., 2008; Grantcharov & Funch-Jensen, 2009; Jenison et al., 2012; Khan et al., 2003; Korndorffer, Dunne, et al., 2005; Lendvay et al.,

2013; Mashaud et al., 2010; Miskovic et al., 2011; Nugent et al., 2013; Rooney et al., 2012; Sanfey et al., 2010; D. J. Scott et al., 2007; Shane et al., 2008; Stack et al., 2010; Stefanidis et al., 2005; Stefanidis et al., 2006; Sutton et al., 2013; van Dongen et al., 2011; Verdaasdonk et al., 2008; Zendejas et al., 2012).

Discussion

Using a systematic search strategy, this review identified the methods by which technical competence is assessed in surgical trainees. The review also identified the validity and reliability of the current assessment methods and whether standards were set to differentiate competent versus non-competent trainees. Furthermore, the quality of evidence was determined using the GRADE classification, in order to provide an idea of the transferability or confidence in the estimate of effect if these methods were employed in other institutions. To date, this is the only review evaluating the methods by which we currently assess technical competence. These findings are important to a variety of stakeholders, including the training programs and governing bodies that are attempting to determine how best to evaluate their competency-based educations paradigms, and trainees for promotion and matriculation.

One interesting phenomenon that emerged in the early stages of the systematic review was the lack of uniformity in the definition of technical competence, the various types of competence that exist and how the literature differentiates competence from proficiency.

The definition of technical competence continues to be debated within the medical literature. There have emerged various terminology domains that incorporate within them an assessment of technical competence (Bhatti & Cummings, 2007; Grober & Jewett, 2006; Hall, Crebbin, & Ellison, 2004; Karamichalis, 2012; Mohr, Batalden, & Barach, 2004; Thomas, 2006; van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). These include; technical competence (Karamichalis, 2012; Mohr et al., 2004), surgical competence (Bhatti & Cummings, 2007; Hall et al., 2004; van der Vleuten et al., 2010), and operative competence (Grober & Jewett, 2006; Thomas, 2006). Multiple publications use these terms interchangeably, however, after further review, it was determined that they have different working definitions (Bhatti & Cummings, 2007; Grober & Jewett, 2006; Hall et al., 2004; Karamichalis, 2012; Mohr et al., 2004; Thomas, 2006; van der Vleuten et al., 2010). Although poorly defined, technical

competence is often used as a synonym for technical skill. One useful explanation comes from Mohr *et al*. who described technical competence as the skill to safely and successfully perform the required steps of a particular procedure, with their research investigating clinical Microsystems (Karamichalis, 2012; Mohr et al., 2004). In the reviewed literature and in other publications, most definitions have embedded within them a minimum standard needed to be safe and/or the ability to perform the operation independently without help (Adrales et al., 2004; A. Ahmed et al., 2013; Bath & Wilson, 2007; Fleming et al., 2012; Francis et al., 2010; Fried et al., 2012; Goff et al., 2005; Gofton et al., 2012; Hawes et al., 1986; Howells et al., 2008; Ishman et al., 2010; Jelovsek et al., 2010; Laeeq et al., 2009; Laeeq et al., 2010; Lin et al., 2009; M. Martin et al., 1998; Miskovic et al., 2011; Pandey et al., 2006; Proctor et al., 1998; Stack et al., 2010; Swanstrom et al., 2006; Taravella et al., 2011; Taylor et al., 2007; Vassiliou et al., 2005). Surgical competence has been described as a collection of skill, knowledge and judgment required to complete new or familiar tasks incorporating both technical and non-technical (i.e. clinical problem solving) components (Bhatti & Cummings, 2007; Hall et al., 2004; van der Vleuten et al., 2010). Furthermore, operative competence builds on this definition but also includes a surgeon's experience and opportunities to form a broader, more universal domain (Grober & Jewett, 2006; Thomas, 2006).

Findings from the reviewed literature also illustrate that there remains confusion when differentiating the terms competence and proficiency. Differentiating these terminologies however, is crucial in the field of education research and an important first step in evaluating methods of assessment.

The Dreyfus and Dreyfus (Dreyfus, 1980) model of skill acquisition differentiates a competent and proficient learner by their situational recognition, with a proficient learner having a more holistic view of a situation and being more advanced than a *competent* learner. Furthermore, competence within the Oxford English Dictionary 2nd edition ("The Oxford English Dictionary," 1989) and New Webster's Dictionary ("Webster's Third New International Dictionary ", 1961) is defined as "sufficiency of qualification" and "answering all requirements, being suitable", both definitions, employ within them a component of legal capacity or power, the ability in essence for completion of training and initiation of independent practice (Gallagher et al., 2005). In the Oxford English Dictionary 2nd edition ("The Oxford English Dictionary," 1989) and New Webster's Dictionary ("Webster's Third New International Dictionary ", 1961),

proficiency has been defined as “advanced in the acquirement of skill” and “well advanced in any branch of knowledge or skill”. In contrast to competence, Gallagher *et al*. defined proficiency as the ability to acquire and consistently perform psychomotor skills, without necessarily being an expert at those skills (Gallagher et al., 2005). Proficiency then, is not merely being sufficient for qualification and independent practice but further advanced, consistent and refined in ones performance.

Given the discrepancy in definitions, the studies included in this review did not have to conform to the definition of technical competence as defined here. Rather, any studies where the investigators described their working definition and subsequently measured this definition were included within the review. For example, if the investigators used the terms surgical competence and technical competence interchangeably but were measuring technical competence, these were included. Moreover, if the investigators were measuring technical proficiency, or technical skill yet continued to refer to it as technical competence, these were included. Conversely, if the investigators were measuring technical competence yet continued to refer to it as technical proficiency, these were also included. In essence, the goal of the review was to capture all the methods by which technical competence, correctly or incorrectly, is currently assessed.

Likert scales as a method of technical competence assessment comprised the largest category in this review. The likert scale assessments included here, all employed the original assessment methods used in the OSATS exam, developed by Reznick *et al*. (Reznick et al., 1997) or a modification of those methods, both of which have been extensively validated and integrated (in part) into multiple residency training programs (K. Ahmed, Miskovic, Darzi, Athanasiou, & Hanna, 2011; Chipman & Schmitz, 2009; Faulkner, Regehr, Martin, & Reznick, 1996; Faurie & Khadra, 2012; Hance et al., 2005; MacRae, Regehr, Leadbetter, & Reznick, 2000; J. A. Martin et al., 1997; Paisley, Baldwin, & Paterson-Brown, 2001; Reznick et al., 1997; D. J. Scott et al., 2000; van Hove, Tuijthof, Verdaasdonk, Stassen, & Dankelman, 2010; Vassiliou et al., 2005). They assessed technical competence by using either the global rating scale, task-specific rating scale or a combination of both. The OSATS (Reznick et al., 1997) exam and in particular its methods of assessment were developed to objectively evaluate the technical skills of trainees. The studies in this review, which employed such assessment methods, for the most part continued to measure technical skill, rather than technical competence, based on the definition provided above. This further exemplifies the need for a unified definition of

technical competence and the need to define and measure a common entity. These studies most often employed a minimum score (i.e. 25/35 for the global rating scale) on their respective rating scales to deem a resident as having passed or being competent on a particular task or procedure. The importance of a minimum score to determine technical competence cannot be understated; however this pass score cannot be arbitrary as is currently common practice (van Hove et al., 2010). Rather, it must be a valid means of differentiating competent versus non-competent trainees. One such practice is to use standard setting methodology adapted from other educational tests (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001). In this review we found only four such studies, three in this category and one in the benchmark category, which attempted to establish standards to differentiate a pass/fail cutoff for technical competence for surgical trainees.

Benchmarking is a method by which current trainees undergoing assessment are compared to previous test takers and their pre-defined levels of performance (Berkey, 1994; Berwick, 1989; Camp, 1992; Hahn et al., 2011; Kiefe et al., 1998; Weissman et al., 1999). These test takers can be fully trained individuals in the field, such as staff surgeons, producing expert benchmarks (Weissman et al., 1999). They can also be other trainees that are non-experts, but the results from whom create relative referenced benchmarks (Brydges et al., 2006; Ganju et al., 2012; von Websky et al., 2012). These predefined levels of performance are then collated regardless of how they are acquired and set as the minimum level required for completion to be deemed competent (Seymour et al., 2002) or some measure away from the predefined scores to be deemed competent (Stefanidis et al., 2010; van Dongen et al., 2011) (i.e. one standard deviation below the level of an expert). The benchmarks are often set using virtual reality training systems or other simulated environments, however they can also be established using other platforms such as recorded or live assessments completed intra-operatively or on standardized bench models (Verdaasdonk et al., 2008). One important consideration within this category lies in how these benchmarks are set. The majority of benchmarking assessments are completed using a virtual reality platform as summarized above. These platforms are often created by the manufacturer with predefined benchmarks for each individual task, these benchmarks however are often arbitrary and not evidence based (Brinkman et al., 2012; Brunner et al., 2005; Grantcharov & Funch-Jensen, 2009; Korndorffer, Scott, et al., 2005; Lendvay et al., 2013; A. K. Moore et al., 2008; Perrenot, 2011; Selvan, 2011; Shane et al., 2008; Stefanidis et

al., 2005; Verdaasdonk et al., 2008). Studies have identified this as a major limitation of these platforms and there is research in creating and validating new benchmarks (Hahn et al., 2011). Breaking down the concept to its core, should benchmarks be at the level of an expert, even though competence does not imply expertise (Gallagher et al., 2005) and are relative referenced benchmarks at an acceptably high level to determine technical competence? Furthermore, should we expect trainees to be as masterful in the operating room as their mentors who have been in practice for an extended period of time? Finally, who or how do we determine that two standard deviations from an expert is sufficient or that to differentiate truly competent versus non-competent performers we should use one standard deviation instead? These questions do not have answers within the realm of surgical education as of yet.

Binary assessments can take on a variety of forms. The studies in this systematic review can be categorized into either binary global competence outcomes or binary task-specific outcomes. The former assesses whether a trainee is competent to perform a procedure overall similar to a global rating scale, using a Yes/No annotation, while the latter assesses whether a trainee satisfies each task that in total makeup a procedure, often using a checklist. These binary assessments, use either a gestalt opinion to make a decision about overall competence as is the case with the global competence outcome, or the parts are greater than the sum opinion as is the case with the task-specific outcome. The downside with this type of assessment is the subjective nature of the global competence outcome. The raters can identify and differentiate performances that are competent or non-competent, but they often cannot explain why or how they came to those conclusions, making this difficult to reproduce with other, perhaps less experienced raters. This has previously been identified in the literature as well, that experts know a qualified trainee when they see one, but that this is overly subjective in nature and they often cannot explain the ways by which they came to this conclusion (Ansell et al., 1979; Maxim & Dielman, 1987; D. Sloan, . Donnelly, M,. Drake, D,. Schwartz, R,. 1993; D. A. Sloan, Donnelly, Schwartz, & Strodel, 1995). While the task-specific competence outcome, like the task-specific scale is procedure-specific, limiting its broad use.

The novel tools utilized in this review were unique in that they attempted to assess technical competence by creating new tools solely for that purpose, instead of modifying existing tools initially created for other purposes including the assessment of technical skill. These novel tools essentially assessed the comfort level of staff surgeons in allowing a trainee to perform a

procedure without their direct influence, by ascertaining whether the staff needed to be present in the room and the degree of their operative involvement. Another method included analyzing errors on intra-operative procedures and inferring technical competence. The major drawback of theses novel tools is the lack of their objectivity and the introduction of resident bias in assessing technical competence. Similar to un-blinded assessments in any study, an individual staff Surgeon's experience with and the reputation of a resident can influence the staff surgeon's comfort with a trainee, dictating the amount that a trainee can and will operate, inferring a false or incorrect level of technical competence. This further perpetuates the traditional methods of informal feedback that are subjective in nature and have been previously shown to be flawed (L. S. Feldman, Hagarty, Ghitulescu, Stanbridge, & Fried, 2004; Wanzel, Ward, & Reznick, 2002).

Surrogate markers to assess technical competence are not a new method to assess surgical performance, previous attempts have been made to assess the performance of certified surgeons, using patient morbidity and mortality as surrogate outcomes (Alaraj et al., 2013; Asimakopoulos et al., 2006; Birkmeyer et al., 2013; Chiu et al., 2006). The major drawback of the assessment methods included here are whether surrogate markers are a true measure of technical competence. There are many other variables at work during a procedure; perhaps the time to complete a procedure, the cecal intubation rate or the strength of tendon repair, are based more on patient characteristics and the peri-procedure environment (team dynamics, bowel preparation) than on resident technical competence. These variables can come into play in all of the categories by which technical competence is assessed, but more so when an overall procedure is not evaluated, although it may be viewed, but rather when a rater focuses solely on the surrogate measures to make an assessment.

One of the major objectives of our study was to not only identify the methods by which technical competence is assessed, but to also evaluate the validity and reliability of the assessment methods. As previously outlined (Dauphinee, Blackmore, Smee, Rothman, & Reznick, 1997) (Roberson, Kentala, & Forbes, 2005), assessment in high-stakes settings, such as resident promotion and matriculation, should meet certain basic criteria for validity and reliability. The studies in this review had their methods validated and their reliability documented in the paper included or in previous studies using the same assessment methods in a different setting. It is important to note the fact that results from previous studies were included only if the assessment method was used without change from the time it was validated and its

reliability was assessed. It is inappropriate to refer to a particular method as having such psychometric properties, when the method is a modification from its original form. Furthermore, the psychometric properties found in one particular study, are not generalizable if the assessment method has been altered in any way.

To assess for transferability of the assessment methods to future studies conducted at other institutions or the degree to which further research was unlikely/likely to change our confidence in the estimate of effect, we assessed the GRADE (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008) classification of each study. The GRADE (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008) classification was assessed for the primary outcome in each study, regardless of what that outcome was. Although the aim of this systematic review was to focus on the assessment methods and their psychometric properties, rather than the primary outcomes of each study, the GRADE (Guyatt, Oxman, Kunz, et al., 2008; Guyatt, Oxman, Vist, et al., 2008) classification assesses the methodological quality of a study and in essence how well the study was carried out and the results reported.

Conclusion

We have provided a unified definition of technical competence based on the published literature. We have also demonstrated that currently there are five methods by which technical competence is assessed and that for the majority of these methods their validity and reliability have previously been examined. Each of the assessment methods have specific drawbacks and most were established to assess technical skill, with a few novel instrument specifically created to assess technical competence. Very few studies have implemented standard setting approaches to differentiate competent and non-competent performers. As we move towards CanMEDS 2015 ("Competency-based medical education," 2013) and the implementation of milestones, it is our view that the best approach to assess technical competence will be to utilize an existing tool in a new environment. We believe that tool is the global rating scale from the initial OSATS examination (Reznick et al., 1997) (likert scale assessment group). The global rating scale (Reznick et al., 1997) is a single tool that is transferable across different trainee levels and surgical procedures, simplifying assessor training and implementation. We caution modification of the global scale (Reznick et al., 1997) however, as it may jeopardize reliability and validity. The new environment is the operating room, not the simulation laboratory, where blinded and

trained raters evaluate entire procedures and not just portions of procedures. Finally, standard setting (J. J. Norcini, 2003; Schindler et al., 2007; T. J. Wilkinson et al., 2001) methodologies should be the focus of further research, whereby for the first time this existing tool and new environment can provide non-arbitrary cut-offs to differentiate competent versus non-competent performers.

1.6.3 Unpublished technical performance assessment practices

Despite the robust information depicted in the systematic review; with the rapid change towards CBME and workplace assessments, there are indisputably technical performance assessments taking place in Canada and internationally for which little documented literature exists. Therefore, this lack of published data regarding current and ideal technical performance assessments goes on to form Aim #2 in this thesis.

1.7 Assessment of non-technical performance

1.7.1 Impetus for assessing non-technical performance

Non-technical skills are defined as a set of social, personal and cognitive abilities that complement a trainee's technical skills and contribute to the overall performance of a given task (R. Flin, O'Connor, P, Crichton, M, 2008; Youngson & Flin, 2010). The skills vary slightly depending on the specific assessment instrument, however most have embedded within them the following constructs: decision making, communication, teamwork, leadership, situational awareness and stress/fatigue management (R. Flin, O'Connor, P, Crichton, M, 2008). Over the last several years multiple studies have demonstrated that the non-technical performance of staff surgeons or trainees has a direct impact on patient outcomes (morbidity and mortality) (Firth-Cozens & Mowbray, 2001; Greenberg et al., 2007; Lingard et al., 2004; Mishra, Catchpole, Dale, & McCulloch, 2008). Specifically Greenberg *et al*. and Lingard *et al*. both demonstrated how communication failures resulted in injury to surgical patients, while Firth-Cozens *et al*. and Mishra *et al*. demonstrated how a lack of leadership and situational awareness respectively, negatively impacted patient care (Firth-Cozens & Mowbray, 2001; Greenberg et al., 2007; Lingard et al., 2004; Mishra et al., 2008). More recently, a review by Boet *et al*. demonstrated that training in some aspects of non-technical performance improved patient outcomes in Surgery, Obstetrics and Gynecology, and Pediatrics (Boet et al., 2014).

These findings coincide with the transition towards CBME in surgical education and an increased focus on these non-technical skills within the CBME frameworks ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; Cogbill, 2014; J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015; "The Good medical practice framework for appraisal and validation," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012). The first non-technical assessment modalities were introduced in the early 2000s by research that originated at the University of Aberdeen, following suit from the field of aviation training, where non-technical performance has been under investigation for much longer (R. Flin, Martin, R., Goeters, K.M., Hormann, H.J., Amalberti, R., Valot, C., Nijhuis, H, 2003).

1.7.2 Instruments for non-technical performance assessment

There are currently four instruments utilized for surgical non-technical performance assessments: Non-Technical Skills (NOTECHS), the Observational Teamwork Assessment for Surgery (OTAS), Non-Technical Skills for Surgeons (NOTSS) and the Objective Structured Assessment of Non-Technical Skills (OSANTS) (Dedy, Szasz, et al., 2015; R. Flin, Martin, R., Goeters, K.M., Hormann, H.J., Amalberti, R., Valot, C., Nijhuis, H, 2003; Healey, Undre, & Vincent, 2004; Yule, Flin, Paterson-Brown, Maran, & Rowley, 2006). The first two are directed at assessing the entire surgical team, while the latter two are directed at assessing individuals, with the OSANTS specifically aimed at assessing trainees.

Regardless of which assessment instrument is utilized, all of them are based on previously established skill taxonomies and behavior marker systems (Yule, Flin, Paterson-Brown, & Maran, 2006). As described by Yule *et al*. and Carthey *et al*. skills taxonomies are the surgical and psychological collection of interpersonal and cognitive skills possessed by surgeons (i.e. communication, teamwork, situational awareness etc.) and behavior marker systems are observable indicators of performance (either individual or team related) embedded within each skills taxonomy (Carthey, 2003; Yule, Flin, Paterson-Brown, & Maran, 2006). Overall, these skills taxonomies and behavior markers are used to organize the training and assessment of non-technical performance, with each non-technical assessment instrument category based on the skills taxonomy and each scale item descriptor (on a Likert scale) or element within each category, based on behavior markers (Yule, Flin, Paterson-Brown, & Maran, 2006). As a result,

non-technical performance instruments cannot only be used for trainee assessment, but also learning.

1.7.3 Non-Technical Skills (NOTECHS)

Developed in the early 2000s through a partnership between the aviation industry and psychologists from the University of Aberdeen, NOTECHS was created as a tool to teach and assess flight crews' non-technical skills (R. Flin, Martin, R., Goeters, K.M., Hormann, H.J., Amalberti, R., Valot, C., Nijhuis, H, 2003; R. Flin, O'Connor, P, Crichton, M, 2008). Given the need for non-technical performance teaching and assessment in surgery as CBME was taking hold and the involvement of Dr. Flin from the University of Aberdeen, a renowned human factors expert, NOTECHS was adapted for use in the OR (R. Flin, O'Connor, P, Crichton, M, 2008; Mishra, Catchpole, & McCulloch, 2009). NOTECHS is composed of four main categories that include: leadership and management, teamwork and cooperation, problem solving and decision-making, and situational awareness (Mishra et al., 2009). Under each of these categories there are between three and five specific elements that help in the assessment, and each category is rated on a four point Likert scale ranging from 1 – below standard to 4 – excellent, for a maximum score of 16 (Mishra et al., 2009). In NOTECHS each OR sub team (surgical, anesthesia, and nursing) is assessed independently as unit (Mishra et al., 2009). Since 2009, the original NOTECHS has been updated to the NOTECHS II to further align it with the specific needs of the OR team and distance it from the aviation industry (E. R. Robertson et al., 2014).

Recently, Morgan *et al*. demonstrated that in Orthopedic Surgery, training interventions based on the NOTECHS II assessment instrument prior to surgery, significantly improved the teamwork performance of surgeons, anesthetist and nurses collectively in the intervention group compared to the control group (Morgan et al., 2015). Furthermore, Robertson *et al*. duplicated these findings using the NOTECHS II assessment instrument as an intensive pre-operative training intervention for Plastic Surgeons and their OR teams, again demonstrating increased non-technical performance in the intervention group compared to the control group (E. Robertson et al., 2015).

1.7.4 Observational Teamwork Assessment for Surgery (OTAS)

Developed at Imperial College London in 2003, OTAS is aimed at assessing overall team tasks and behaviors that contribute to performance as it relates to patient safety in the OR (Healey et al., 2004). There are five main categories that make up OTAS and they include: communication, co-operation, co-ordination, leadership, and monitoring (Healey et al., 2004). Each of these categories is assessed on a seven point Likert scale ranging from 0 – team function non-existent to 6 – enhanced team function, for a total score of 35 (Healey et al., 2004). Unlike NOTECHS above, the entire OR team is assessed as a unit, with no sub team breakdown (Healey et al., 2004).

Recently, Phitayakorn *et al*. used a variety of teamwork assessment instruments including OTAS to evaluate a team of surgical and anesthesia residents working in a simulated OR alongside a group of nurses (Phitayakorn, Minehart, Hemingway, Pian-Smith, & Petrusa, 2015). The purpose of the study was to determine whether teamwork as assessed using OTAS correlated with the team's ability to manage a medical emergency (malignant hyperthermia) by following a specified procedural checklist for its treatment (Phitayakorn et al., 2015). They documented that although the OTAS instrument was an accurate assessment of team performance when used by various assessors, there was almost no correlation between OTAS scores and the ability to complete the procedural checklist, consequently suggesting training regimens to better align non-technical performance and the management of medical emergencies (Phitayakorn et al., 2015).

1.7.5 Non-Technical Skills for Surgeons (NOTSS)

Developed in 2006 at the University of Aberdeen, NOTSS is geared at providing training and feedback in non-technical performance, aimed specifically at practicing surgeons, although it has also been adapted for use with residents (Yule, Flin, Paterson-Brown, Maran, et al., 2006). There are four main categories that make up the NOTSS assessment and they include: situational awareness, decision-making, communication, and teamwork and leadership (Yule, Flin, Paterson-Brown, Maran, et al., 2006; Yule et al., 2009). Under each of these categories there a three elements that help in completing the assessment, and each of the main categories are rated on a four point Likert scale from 1– poor to 4 – good, to give a total score out of 16 (Yule, Flin, Paterson-Brown, Maran, et al., 2006; Yule et al., 2009).

Since it's inception, NOTSS has been used in surgery across a variety of spectrums (Pena et al., 2015; Yule et al., 2015). Yule *et al*. in 2015 demonstrated that trainees who received non-technical feedback based on the NOTSS assessment instrument from a coach incrementally over a span of five straight operations, significantly improved their non-technical performance during simulated laparoscopic cholecystectomies compared to control trainees (Yule et al., 2015). Similarly, Pena *et al*. in 2015 completed a study where they randomized and trained residents' non-technical skills to varying degrees (a simulation plus didactic session or simulation session alone) and submitted them to three simulated crisis scenarios dispersed across six weeks (Pena et al., 2015). They found that regardless of the group, exposure to NOTSS training improved their perception of how such non-technical performance will impact their future practice and improved their no-technical performance compared to their baseline performance (Pena et al., 2015).

1.7.6 Objective Structured Assessment of Non-Technical Skills (OSANTS)

Developed at the University of Toronto in 2014, OSANTS provides training and feedback in non-technical performance geared specifically at surgical trainees (Dedy, Szasz, et al., 2015). There are seven main categories and they include: situational awareness, decision-making, teamwork, communication, leading and directing, professionalism, and managing and coordinating (Dedy, Szasz, et al., 2015). Each of these categories is ranked on a global five point Likert scale with specific descriptions/explanations for each item for a composite score out of 35 (Dedy, Szasz, et al., 2015). OSANTS addresses the shortcomings present in the other assessment modalities that were really developed for staff surgeons and adapted for use with trainees after the fact (Dedy, Szasz, et al., 2015). In particular, it focuses on two domains (leading and directing, and managing and coordinating) that trainees usually lack experience with, compared to staff surgeons where a lack of such non-technical ability does not usually influence staff performance (Napolitano et al., 2014).

A recent study by Dedy *et al*. used the OSANTS assessment instrument as a method of feedback for surgical trainees that took part in a study in the real OR (Dedy, Fecso, et al., 2015). Trainee participants were evaluated in the OR across a variety of surgical procedures and their non-technical performance was assessed at baseline and at a subsequent surgical procedure following a debriefing session using the OSANTS instrument (Dedy, Fecso, et al., 2015). The

use of the OSANTS assessment instrument significantly improved the average non-technical scores in trainees, from their baseline to post-intervention performance (Dedy, Fecso, et al., 2015).

1.7.7 Other trainee factors which may contribute to overall performance

There are ostensibly other intangible qualities on top of the technical and non-technical skill possessed by trainees, which too may contribute to the performance of a particular task or procedure (R. M. Bell, Fann, Morrison, & Lisk, 2011). Although these abilities are beyond the scope of this research and each of them complex in their own right, their acknowledgement is warranted at this moment. Our focus in this thesis however, was solely on trainee technical and non-technical performance.

1.8 Performance assessors

1.8.1 Types of assessors

Up to this point, the types of assessment categories have been discussed (formative versus summative) as well as the methods by which technical and non-technical performance are assessed. The use of Messick's conceptual framework of validity was also introduced to ensure that the assessment results utilized are valid (and reliable), focusing in particular on the need for assessor training (D. A. Cook & Beckman, 2006; M. Feldman et al., 2012; Messick, 1995). Recently there has been some discourse however, in regards to who the assessors ought to be during both formative and summative assessments of technical and non-technical performance (Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007; Hawkins et al., 2015; Holmboe et al., 2010; Schuwirth & Van der Vleuten, 2011; Swing et al., 2009; T. J. Wilkinson & Wade, 2007). Specifically, whether the assessors need to be removed from the training process (external assessors) or whether those who train residents, can also be involved in assessing them (internal assessors) – particularly for summative assessments (Govaerts et al., 2007; Hawkins et al., 2015; T. J. Wilkinson & Wade, 2007).

1.8.2 External assessors

External assessors are individuals outside of the organization or program taking part in the assessment, with no vested interest in how those within the organization fair (Chen, 2015a). In the context of PGME, trainees within the residency program are the individuals being

assessed. The external assessors often have no relationship with the program and no relationship with the trainees, as such they are said to have independence (Chen, 2015a). An example of an external assessor in General Surgery is a visiting researcher, who is also a surgeon, assessing trainees in the OR as part of a large multi-institution research project. Another example of using external assessors in General Surgery is the technical skills OSCE undertaken by all surgery residents (including those in General Surgery) at the University of Toronto (S. de Montbrun, Satterthwaite, & Grantcharov, 2016). At the completion of the PGY1 year, surgical residents take an examination focused on the basic skills required of all surgical trainees (i.e. chest tube insertion, tracheostomy, skin incision closure etc.) (S. de Montbrun, Satterthwaite, et al., 2016). The assessors although faculty at the University of Toronto, come from a variety of backgrounds, do not know the trainees and given there is no consequence to the faculty from this assessment, have no stake in the outcomes (S. de Montbrun, Satterthwaite, et al., 2016).

1.8.3 Internal assessors

Internal assessors are individuals who are members of the organization or program taking part in the assessment (Chen, 2015a). They are familiar with the program and often have a stake in the outcome of the evaluation (Chen, 2015a). These internal assessors often have a relationship with the program, individual trainees or both, and as such they are said to have limited or no independence (Chen, 2015a). An example of an internal assessor in General Surgery is a staff surgeon on faculty who a trainee has been working with over an extended period of time in variety of settings (i.e. outpatient clinics, the OR, tumor boards etc.).

1.8.4 Utility of internal assessors

The motivation for utilizing internal assessors mainly focuses on their ability to address and minimize the many issues inherent to using external assessors (Chen, 2015a; Takahashi, 2011). These issues include the logistics of attaining and familiarizing external assessors with the specific programs and trainee facets they are to assess and the sizeable cost required to develop, implement and maintain the use of external assessors as part of the assessment framework (Chen, 2015a; Takahashi, 2011).

The motivation against using internal assessors for assessment really focuses on the risk of bias they may incorporate into the assessment framework (Norman, Van der Vleuten, & De

Graaff, 1991). This bias can be a result of the staff surgeon's previous knowledge of a trainee (either direct or indirect), their previous experience with a trainee in a variety of settings, and an admiration for the trainee for a multitude of reasons (i.e. they remind the staff surgeon of themselves) (Reid, Kim, Mandel, Smith, & Bansal, 2014; T. J. Wilkinson & Wade, 2007; Williams, Klamen, & McGaghie, 2003). Finally, if a staff surgeon has worked with a trainee closely over many months, there is presumed to be a conflict of interest between teaching and assessing (they are confounded) – if the assessor is responsible for training a resident, they are not only assessing resident performance, but as a result, also their ability to teach (T. J. Wilkinson & Wade, 2007).

1.8.5 Comparison between external and internal assessors

There is a paucity of evidence directly comparing external and internal assessors in terms of their accuracy or reliability. The literature that is available compares internal assessors with more objective assessments (i.e. written examinations, structured technical skill simulation examinations) or less likely biased assessors (i.e. resident self or peer assessment), yet these individuals are still clearly internal to the overall assessment process (Elfenbein et al., 2015; Farrell et al., 2010; L. S. Feldman et al., 2004; Ray et al., 2016; Reid et al., 2014; Steigerwald, Park, Hardy, Gillman, & Vergis, 2015). The available studies then in reality compare internal assessors with surrogate performance outcomes. Nonetheless, they do provide some data in an area where evidence is currently scarce.

Studies comparing internal assessors (staff surgeons or staff physicians) with multiple choice examinations have been completed for varying levels of trainees, including medical students and residents, revealing mixed results (Elfenbein et al., 2015; Farrell et al., 2010; L. S. Feldman et al., 2004; Ray et al., 2016; Reid et al., 2014; Steigerwald et al., 2015). Farrell *et al*. and Reid *et al*. compared the overall clinical performance of medical students as assessed by internal assessors to their performance on the National Board of Medical Examiners (NBME) end of rotation examination, demonstrating a weak and moderate correlation, respectively (Farrell et al., 2010; Reid et al., 2014). Specifically at the PGME level, Ray *et al*. and Elfenbein *et al*. recently demonstrated that there were virtually no correlations between General Surgery residents' clinical rotation evaluations completed by internal assessors and passing the annual ABSITE examination (Elfenbein et al., 2015; Ray et al., 2016). This work in the area of PGME

is further supported by the pioneering work initially completed by Feldman *et al*. and more recently Steigerwald *et al*., comparing internal assessors to standardized technical skill examinations (L. S. Feldman et al., 2004; Steigerwald et al., 2015). In the study by Feldman *et al*. they compared ITERs completed by internal assessors to a technical skills examination of basic laparoscopic tasks, demonstrating a weak correlation between the two, but an inability of the ITER to discriminate between lower performing trainees (L. S. Feldman et al., 2004). In the study by Steigerwald *et al.*, they compared OR evaluations completed by internal assessors of trainees performing a laparoscopic cholecystectomy to a technical skills examination of both basic laparoscopic skills and a virtual reality basic task, demonstrating no correlation between the two (Steigerwald et al., 2015).

Studies comparing various levels of internal assessors have also been completed for medical students and residents, also revealing mixed results (Goldstein et al., 2014; Herrera-Almario et al., 2016; Moonen-van Loon et al., 2015; Ward et al., 2003). Goldstein *et al*. compared the clinical knowledge ratings of medical students, completed by faculty and residents (internal assessors) the medical student had worked with, showing a reasonable correlation within each group, but moderate (20%) variance between the groups (Goldstein et al., 2014). At the PGME level, Moonen-Van Loon *et al*. compared resident evaluations completed by the medical faculty, nurses, administrative personnel and allied health team members (all internal assessors) demonstrating good reliability when multiple raters watched trainees on multiple occasions, yet the internal assessors contributed an unknown, but likely substantial amount of variance to the overall score, as 80% of total variance could not be accounted for in this study design – including the variance among raters (Moonen-van Loon et al., 2015). This work in the realm of PGME is further supported by Ward *et al*. and Herrera-Almario *et al*. more recently, comparing faculty and resident self-assessments (Herrera-Almario et al., 2016; Ward et al., 2003). The study by Ward *et al*. compared faculty and resident self assessments (both internal assessors) for a laparoscopic Nissen fundoplication, showing a weak to moderate correlation between the two assessor groups (Ward et al., 2003). While the study by Herrera-Almario *et al*. again compared faculty and resident self assessment (internal assessors) for a variety of laparoscopic procedures, and although no correlations were provided, the assessments completed by the residents were statistically different (lower) for almost every operative task of each procedure compared to the faculty assessments (Herrera-Almario et al., 2016).

Despite this work, most of it very recent, no studies have compared internal and external assessors in the OR. As a result there is no information about whether internal raters once trained, can be utilized in surgical education for performance assessments that are either summative or formative. Determining whether there is an association between internal and external assessors for both technical and non-technical performance assessments in the OR forms the basis of Aim #4 in this thesis.

1.9 Performance standards

1.9.1 Basis of performance standards

Once the appropriate methods, and assessors for evaluating technical and non-technical performance have been selected, the next step is to ensure the decisions stemming from such assessments are based on established performance standards.

Standard setting is a process of generating cut scores on examinations or assessments in a systematic and methodical manner (Cizek, 1993; Cizek & Bunch, 2007f; J. J. Norcini, 2003). As stated by Cizek *et al*., standard setting is "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance" (Cizek, 1993). These cut scores (standards) then divide the performance of trainees into two states or levels for a particular task; those that have attained the standard (i.e. are competent) and those that have not (i.e. are non-competent) (Cizek & Bunch, 2007f; J. J. Norcini, 2003; J. J. Norcini, Shea, JA, 1997). This is in contrast to what is currently done in much of education, whereby performance metrics (scores needed to be deemed competent for example) are chosen arbitrarily (Cizek & Bunch, 2007c). One important concept is to draw a distinction between the cut score and the performance standard, although they are often used interchangeably (Cizek & Bunch, 2007f). The cut score, alternatively called the passing score, is a point on a scale (i.e. OSATS scale), while the performance standard is the minimal level of preferred performance (i.e. competence) (Kane, 1994). As stated by Kane, "the performance standard is the conceptual version of the desired level of competence, and the passing score is the operational version" (Kane, 1994).

1.9.2 Types of standards

Standards can be classified as either criterion-referenced, which have been shown superior (and are almost exclusively utilized, and solely discussed here) for all assessments including those that are summative in nature - where a cut score is determined and any trainees that reach that score deemed as competent (Schindler et al., 2007; T. J. Wilkinson et al., 2001). Conversely, relative-referenced standards determine a cut score based on the top percentage of trainees assessed (i.e. top 25%) (Schindler et al., 2007; T. J. Wilkinson et al., 2001).

Another important differentiation noted by Jaeger suggests that standards on top of being either criterion or relative-referenced can also be categorized as either test-centered or examinee-centered (Cizek & Bunch, 2007d; Jaeger, 1989). Test-centered strategies set cut scores based on test content and are used predominantly for written examinations (i.e. MCCQE and USMLE examinations), while examinee-centered strategies set cut scores based on the actual performance of trainees (Cizek & Bunch, 2007d; Cohen, Barsuk, McGaghie, & Wayne, 2013; Jaeger, 1989; Nungester, Dillon, Swanson, Orr, & Powell, 1991; Rothman, Blackmore, Dauphinee, & Reznick, 1997; Tolsgaard et al., 2014).

1.9.3 The importance of experts in setting standards

Central to all standard setting is the use of experts (standard setters) (Livingtson, 1982). Although a systematic and prescribed manner is used to arrive at the cut scores, they are based on the judgment of a group of experts and as result the appropriate selection of these participants for a particular purpose is crucial (S. M. Downing, Tekian, A, Yudkowsky, R, 2006). Although there is no real definition of what an expert entails there are a few common characteristics (Cizek & Bunch, 2007b; Livingtson, 1982; J. J. Norcini, Shea, JA, 1997). Firstly, the experts must have the appropriate qualifications and they must be familiar with the types of trainees to be assessed (i.e. Experts in General Surgery should be involved in setting standards for General Surgery procedures and residents) (Cizek & Bunch, 2007b; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Livingtson, 1982; J. J. Norcini, Shea, JA, 1997). Secondly, the experts must understand what standard setting is and that their knowledge and judgment are being utilized to set standards (Cizek & Bunch, 2007b; Livingtson, 1982). Thirdly, the experts must understand the purpose of the examination or assessment for which they are setting standards (i.e. how high-stakes is the summative assessment) (Cizek & Bunch, 2007b; Livingtson, 1982). Finally, the experts must be

trained in the process of setting standards using whichever methodology they are to employ and multiple experts should be used from varying backgrounds (albeit with the appropriate qualifications) (Cizek & Bunch, 2007b; J. J. Norcini, Shea, JA, 1997).

1.9.4 Test-centered standards

Examples of test-centered methodologies include the Angoff and the Hofstee (S. M. Downing, Tekian, A, Yudkowsky, R, 2006).

The Angoff method was first described in 1971 (Cizek & Bunch, 2007a). It is the most common standard setting methodology used for written examinations (Cizek & Bunch, 2007a). In the Angoff method each test question (for a written examination) or scale/checklist item (for performance assessments) is evaluated by standard setting experts (Cizek & Bunch, 2007a; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). These experts decide for each test question or scale item the likelihood that a hypothetical borderline trainee (a trainee who has a 50% chance of passing) would answer the question or perform the task correctly (Cizek & Bunch, 2007a; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). The expert's responses are averaged per question or scale item and the standard is set at the summation of the expert averages (Cizek & Bunch, 2007a; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003).

The Hofstee method was first described in 1983 (Cizek & Bunch, 2007e). In the Hofstee method, experts based on previously accrued evaluation or performance data from a similar population of trainees (i.e. means, standard deviations, quartiles etc.) define the maximum and minimum cut scores and failure rates on a written examination or performance assessment they deem to be acceptable for their set of hypothetical trainees (Cizek & Bunch, 2007e; S. M. Downing, Tekian, A, Yudkowsky, R, 2006). These values are then averaged over all the experts and the resultant numbers for the cut scores and failure rates are subsequently graphed and the intersection between the midpoint of these becomes the standard (Figure 3) (Cizek & Bunch, 2007e; S. M. Downing, Tekian, A, Yudkowsky, R, 2006). Rather than the question or item specific approach taken with the Angoff, the Hofstee uses a more global approach (S. M. Downing, Tekian, A, Yudkowsky, R, 2006).

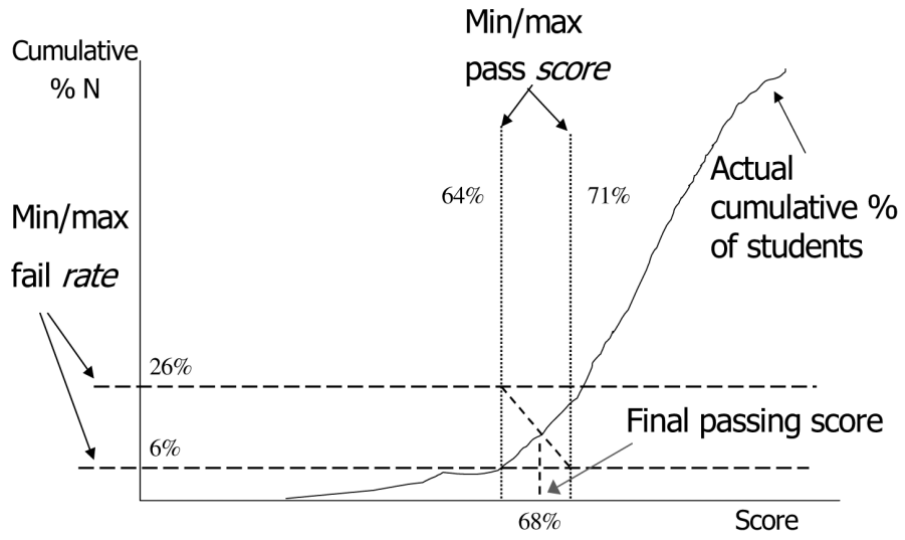

Figure 3. Example of the Hofstee method. The minimum/maximum pass scores are 64% and 71%, respectively while the minimum/maximum fail rates are 6% and 26%, respectively. The performance standard is set at the intersection of their midpoints. Reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com) from Downing *et al*., RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education, *Teaching and Learning in Medicine: An International Journal* (S. M. Downing, Tekian, A, Yudkowsky, R, 2006).

1.9.5 Examinee-centered standards

Examples of examine-centered methodologies include the borderline and contrasting groups (S. M. Downing, Tekian, A, Yudkowsky, R, 2006). In these two methods, experts make decisions not based on the content of an examination or scale/checklist item, or the performance of a hypothetical trainee, but rather directly by watching real trainee performances (Cizek & Bunch, 2007d). These methods bring together overall judgments about trainees and the trainees actual performance to develop standards (Cizek & Bunch, 2007d).

The borderline groups method was first described in 1977 (Cizek & Bunch, 2007d). In the borderline method, experts observe trainees' performing a task or procedure and overall the trainees' are judged to exhibit a pass, fail or borderline performance (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006). Subsequently, either the same expert or another set of experts actual assesses that specific task or procedure using a scale/checklist item to arrive at an actual performance score (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky,

R, 2006). The trainees who overall were deemed to have a borderline performance have their actual performance scores graphed and the median or mean of these actual scores is then set as the performance standard (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006).

The contrasting groups method was first described in 1976 (Cizek & Bunch, 2007d). In the contrasting method, like the borderline method, experts observe trainees performing a task or procedure and overall the trainees are judged to exhibit either a pass or fail (competent/non-competent) performance (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006). Subsequently, again either the same or alternate experts assess the task or procedure using a scale/checklist item to arrive at an actual performance score (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006). The actual performance scores of the trainees in each of the two groups (pass and fail) are graphed (giving rise to normal or near normal distribution with enough observations) and the means and standard deviations are also calculated (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). The performance standard is then set at the intersection of these two (pass and fail) distributions (Figure 4) (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003).

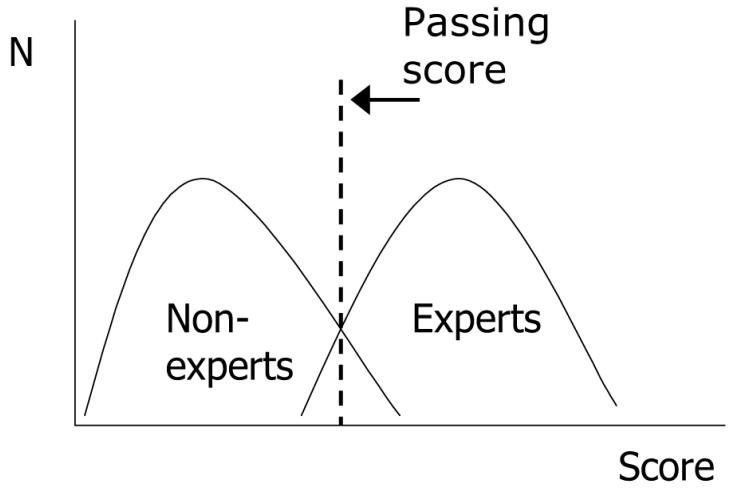

Figure 4. Example of the contrasting groups method. The performance standard is set at the intersection between the pass and fail (non-expert/expert) distributions. The standard can be shifted to the right in order to minimize false positive decisions or to the left to minimize false negative decisions. Reprinted by permission of Taylor & Francis LLC

(http://www.tandfonline.com) from Downing *et al.*, RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education, *Teaching and Learning in Medicine: An International Journal* (S. M. Downing, Tekian, A, Yudkowsky, R, 2006).

The major benefit of using the examinee-centered standards such as the borderline group or contrasting group methods in assessing actual trainee performance is the ability to adjust the performance standard based on the purpose of the assessment (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). If for example the assessment is low-stakes summative in nature, the risk of false negative decisions (trainees that are competent but did no achieve the standard) and false positive decisions (trainees that meet the standard but are in fact non-competent) are of equal weight (Cizek & Bunch, 2007d). If however the assessment is high-stakes summative in nature, the purpose of the assessment will determine whether it is best to minimize the false negative or false positive decisions (S. M. Downing, Tekian, A, Yudkowsky, R, 2006). Take for example a summative assessment at the beginning of training, where there will be multiple future summative assessment opportunities; at this juncture it would be best to minimize false negative decisions (as these can be very costly to the training program and the institutional PGME department) and therefore move the performance standard to the left (Figure 4). Conversely, if this is the last summative assessment a trainee will undertake (i.e. licensing examination) it would be best for the trainee, training program and public to minimize the false positive decisions and therefore move the performance standard to the right (Figure 4). The movement of the performance standard to the left or right is based on tweaking the sensitivity and specificity of the true positive and true negative rates – a concept based in statistical decision theory (Livingtson, 1982).

The major benefit of using the contrasting groups method over the borderline method is based on performance standard accuracy (Cizek & Bunch, 2007d). In order to obtain an accurate standard using the borderline group method, the sample size required to obtain enough borderline trainees (most trainees in education are viewed as either pass or fail) is often too substantial and unfeasible in studies involving direct trainee observations; leading to performance standards that are unstable as a result (Cizek & Bunch, 2007d). Moreover, in the borderline method, by using the mean or median of the actual performance scores of all borderline trainees in creating the performance standard, risks a wide dispersion of scores in those candidates deemed to be

borderline (as has been reported), which may ultimately bias the created performance standard (Cizek & Bunch, 2007d).

1.9.6 Performance standards in medical education

Although test-centered standards have been well established for quite some time in the form of written examinations for various education fields, including medicine, the use of standards (either test or examinee-centered) for performance assessments has only recently gained mainstream attention in the medical community (Holmboe et al., 2010; Holmboe et al., 2011; Szasz, Louridas, Harris, Aggarwal, & Grantcharov, 2015).

Up to this current juncture, three standard setting methodologies have shown to be useful in medicine and to some extent surgery for the performance assessment of trainees in the clinical context and they include: the Angoff, borderline and contrasting groups approaches. Each method's use in medical education is discussed below.

1.9.7 Examples of standards utilized in medical education

Angoff

Studies have been completed using the Angoff method to set standards for technical performance (Barsuk et al., 2012; Cohen et al., 2013; Huang et al., 2009; Jelovsek et al., 2010; Teitelbaum et al., 2014; Walzak et al., 2015; Wayne, Barsuk, O'Leary, Fudala, & McGaghie, 2008; Yudkowsky, Tumuluru, Casey, Herlich, & Ledonne, 2014). The majority utilize internal medicine residents (with one study evaluating General Surgery trainees), in the simulated setting, not the workplace where the procedure normally takes place (and in essence forms the basis of CBME) (Barsuk et al., 2012; Cohen et al., 2013; Huang et al., 2009; Teitelbaum et al., 2014; Walzak et al., 2015; Wayne et al., 2008). Very few (two) studies have assessed trainees directly in the clinical environment (Cohen et al., 2013; Jelovsek et al., 2010). None of the studies, which occurred in the clinical environment, assessed General Surgery trainees. No studies have assessed trainees' non-technical performance (incorporating the other CBME competencies/domains).

Specific examples include, Yudkowsky *et al*. and Walzak *et al*. who set performance standards for medical students and internal medicine residents respectively, using the Angoff

method for a myriad of basic procedural tasks using a variety of experts [medical residents (internal medicine, emergency medicine and cardiology), staff physicians, nurse practitioners and paramedics] (Walzak et al., 2015; Yudkowsky et al., 2014). While Teitelbaum *et al*. set performance standards for senior General Surgery residents performing simulated laparoscopic common bile duct explorations, using the Angoff method with two expert surgeons (Teitelbaum et al., 2014).

Borderline groups

A more limited number of studies have been completed using the borderline groups method to set standards for technical performance (S. de Montbrun, Roberts, Satterthwaite, & MacRae, 2016; S. de Montbrun, Satterthwaite, et al., 2016; Dwyer et al., 2016; Kissin et al., 2013). These studies are split in their use of internal medicine residents and various surgical residents (including one study with General Surgery trainees) (S. de Montbrun, Roberts, et al., 2016; S. de Montbrun, Satterthwaite, et al., 2016; Dwyer et al., 2016; Kissin et al., 2013). No studies have assessed trainees directly in the clinical environment and none have assessed trainees' non-technical performance.

Specific examples include, Kissin *et al*. who set performance standards for Rheumatology residents, using the borderline group method for a collection of eight musculoskeletal joint ultrasound tasks using two Rheumatology experts (Kissin et al., 2013). Dwyer *et al*. who set performance standards for Orthopedic Surgery residents, utilizing a variety of approaches including the borderline groups method for a sports medicine OSCE using various groups of experts (sports surgeons, non-sport surgeons, orthopedic fellows) (Dwyer et al., 2016). While de Montbrun *et al*. in two separate studies set performance standards for General Surgery and Colorectal Surgery trainees, utilizing a variety of approaches including the borderline groups method for a multi-station OSCE for basic (General Surgery trainees) and complex (Colorectal trainees) procedural tasks using a large group of surgeon experts (S. de Montbrun, Roberts, et al., 2016; S. de Montbrun, Satterthwaite, et al., 2016).

Contrasting groups

Several studies have been completed using the contrasting groups method to set standards for technical performance (J. D. Beard et al., 2005; S. de Montbrun, Satterthwaite, et al., 2016;

Diwadkar, van den Bogert, Barber, & Jelovsek, 2009; Jacobsen, Andersen, Hansen, & Konge, 2015; Konge et al., 2013; Konge et al., 2012; Preisler, Svendsen, Nerup, Svendsen, & Konge, 2015; Sedlack, 2011; Sedlack, Coyle, & Group, 2016; Thinggaard, Bjerrum, Strandbygaard, Gogenur, & Konge, 2015; Tolsgaard et al., 2014). These studies are split in their use of various trainees, with two directly assessing General Surgery residents (S. de Montbrun, Satterthwaite, et al., 2016; Thinggaard et al., 2015). Only five studies to date have assessed trainees in the clinical environment; evaluating Gastroenterology trainees performing colonoscopies (two), Obstetric and Gynecology trainees performing either transvaginal/transabdominal ultrasounds or vaginal hysterectomies, and one study evaluating Vascular Surgery trainees completing a disconnection procedure (J. D. Beard et al., 2005; Diwadkar et al., 2009; Sedlack, 2011; Sedlack et al., 2016; Tolsgaard et al., 2014). None of the studies, which occurred in the clinical environment, assessed General Surgery trainees. No studies have assessed trainees' non-technical performance.

For example, more than 10 years ago Beard *et al*. set performance standards for Vascular Surgery trainees using the contrasting group method for a saphenofemoral disconnection performed in the operating room using a multitude of expert raters (composed of Vascular surgeons and residents) (J. D. Beard et al., 2005). More recently, Sedlack *et al*. set performance standards for Gastroenterology trainees using the contrasting group method for colonoscopies performed in the endoscopy suite using multiple expert Gastroenterologists (Sedlack et al., 2016). While Thinggaard *et al*., set performance standards for various residents and staff physicians, including General Surgery trainees using the contrasting groups method for basic laparoscopic skill tasks using expert surgeon raters (Thinggaard et al., 2015).

1.9.8 Performance standard void in surgical education

As can be seen, although three main examinee-centered methods have been utilized to set performance standards in medical education, the use of standards in surgical education is more sparse. While their use to evaluate General Surgery trainees is almost non-existent, particularly in the clinical setting (OR). Furthermore, there is a complete lack of standards in medical or surgical education for non-technical performance. Finally, although there has been some work as reviewed by Hull *et al*. evaluating the association between technical and non-technical performance – demonstrating mixed results (the included studies for the most part assessed performance in the simulated setting with technical and non-technical assessment instruments

not in in keeping with the conceptual framework of validity) (Hull et al., 2012). No studies have assessed the ability of trainees to achieve the technical and non-technical performance standards concurrently. As a result, this lack of technical and non-technical performance standards for General Surgery trainees in the OR and lack of concurrent achievement evidence, goes on to form the basis of Aim #5 in this thesis.

CHAPTER 2

RESEARCH HYPOTHESES AND AIMS

Chapter purpose

This chapter provides the rationale for the overall thesis in the form of a purpose statement. It also helps frame the thesis with major and minor hypotheses, specific research aims (studies contained within the thesis) and their associated objectives.

2 Research Hypotheses and Aims

2.1 Purpose statement

The purpose of this thesis is to collect evidence that will help operationalize a CBME and assessment framework in General Surgery.

2.2 Hypotheses

Main hypothesis

Credible and reliable standards can be set to differentiate between competent and non-competent General Surgery trainees for both technical and non-technical performance during a laparoscopic cholecystectomy in the real OR.

Secondary hypotheses

Trained internal (staff surgeons) raters will provide higher assessment scores for technical and non-technical performance of General Surgery trainees, compared to their trained external (independent to the training process) rater counterparts.

Utilizing a consensus-based technique a contemporary training and assessment model can be developed that begins to form the basis of CBME and assessment in General Surgery.

Canadian training paradigms can be compared to international training paradigms to establish best practice protocols for the implementation of technical performance assessments along the continuum of surgical training, focusing on the in-training time point.

Employing a systematic review strategy a unified definition of technical competence can be established, as can the methods by which technical competence is currently assessed in surgical trainees and whether performance standards have previously been utilized in surgery.

2.3 Aims

The overall aims of this thesis are presented along with their subsidiary objectives directly below.

Aim 1: To systematically review the literature evaluating how technical competence is currently assessed in surgical trainees:

I. Examine the methods by which technical competence is assessed in surgical trainees;

II. Evaluate the validity and reliability of these methods and document the standard setting strategies utilized, if any;

III. Assess the quality of evidence of the included studies; and

IV. Define technical competence and establish how the literature differentiates this from proficiency.

Aim 2: To delineate an international perspective regarding technical competence assessments during all stages of surgical training (selection into training, in-training progression and certification):

I. Identify current technical performance assessment practices employed internationally;

II. Outline ideal technical performance assessment practices (including evaluation instruments and locations of the assessments); and

III. Present real and perceived barriers to the adoption of technical performance assessments.

Aim 3: To identify a contemporary training and assessment model that goes on to form the foundation of CBME in General Surgery:

I. Outline a training model on the operative procedures and tasks that are appropriate for junior and senior level trainees; and

II. Outline an assessment model on the procedures that can be used for technical milestone evaluations, in order to determine whether junior level trainees can progress to senior level trainees.

Aim 4: To determine whether staff surgeons who are intimately tied to training residents, can provide accurate performance assessments:

I. Compare the ratings given be internal (staff surgeons) raters versus external raters (no relationship to the residents) in terms of the technical and non-technical performance scores they attribute to trainees in the operating room during a laparoscopic cholecystectomy.

Aim 5: To produce performance standards that can delineate competent trainees from non-competent trainees:

I. Create a technical and non-technical performance standard for the laparoscopic cholecystectomy;

II. Assess the classification accuracy (reliability) of these newly created performance standards;

III. Assess the credibility (validity) of these newly created performance standards;

IV. Determine a trainees' ability to meet the two performance standards concurrently; and

V. Determine which trainee factors predict standard acquisition.

CHAPTER 3

INTERNATIONAL ASSESSMENT PRACTICES ALONG THE CONTINUUM OF SURGICAL TRAINING

Chapter purpose

This chapter provides an international stakeholder opinion (including that of Canada) regarding the current and ideal technical assessment methodologies employed in various surgical jurisdictions worldwide and the barriers to their ultimate incorporation within a CBME framework. This chapter addresses the limitations identified in Section 1.6.8 of Chapter 1.

Chapter preface

The contents of this chapter have been published in the *American Journal of Surgery* and reprinted with permission from Elsevier Ltd © as: International Assessment Practices Along the Continuum of Surgical Training. M. Louridas*, **P. Szasz***, S. de Montbrun, K.A. Harris, T.P. Grantcharov (2016) *Am J Surg* (E pub ahead of print) DOI: 10.1016/j.amjsurg.2015.12.017. * These authors share first co-authorship

3 International Assessment Practices Along the Continuum of Surgical Training

3.1 Abstract

Background: The objectives of this study were to assemble an international perspective on (1) current and (2) ideal technical performance assessment methods, and (3) barriers to their adoption during: selection, in-training and certification.

Methods: A questionnaire was distributed to international educational directorates (EDs).

Results: Eight of 10 jurisdictions responded. Currently, aptitude tests or simulated tasks (3) are used during selection, observational rating scales (8) during training and nothing is utilized at certification. Ideally, innate ability should be determined during selection, in-training evaluation reports (6) and global rating scales (6) used during training, while global and procedure specific rating scales (6) used at the time of certification. Barriers include lack of predictive evidence for use in selection (5), financial limitations during training (4) and a combination with respect to certification (3)

Conclusion: Identifying current and ideal evaluation methods will prove beneficial to ensure the best assessments of technical performance are chosen for each training time point.

3.2 Introduction

There are several differences in the structure, duration and specific training guidelines between surgical programs internationally ("CanMEDS 2005 Framework," 2005; Cogbill, 2014; "Intercollegiate Surgical Curriculum Overview," 2013). Despite these differences, for the first time in history surgical training has seen an international shift towards a common training paradigm, namely competency-based education (J. R. Frank, Snell, L.S., Sherbino, J, 2014; "Milestones," 2014; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014).

Competency-based education places less emphasis on the duration of training and more on the acquisition and demonstration by trainees of specific competencies – "observable abilities of a health professional "(J. R. Frank, Snell, et al., 2010). These competencies span a variety of domains and can broadly be categorized into: medical expertise, technical performance (on its own or as a subheading under medical expertise), scholarship, professionalism, communication, collaboration within a team setting, patient advocacy and healthcare management ("CanMEDS 2005 Framework," 2005; Cogbill, 2014; "Intercollegiate Surgical Curriculum Overview," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012). Although each of these competencies are extremely important and methods on how best to assess each one are required to implement an all encompassing competency - based assessment framework. The one that differentiates a surgical specialty from a medical specialty and underpins all surgical training programs internationally is technical performance ("CanMEDS 2005 Framework," 2005; Cogbill, 2014; "Intercollegiate Surgical Curriculum Overview," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012). Therefore this was determined to be a good starting point to begin an international collaboration on competency-based education ("CanMEDS 2005 Framework," 2005; Cogbill, 2014; "Intercollegiate Surgical Curriculum Overview," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012). Surgical trainees regardless of specialty are required to attain, and then demonstrate, appropriate and safe operative techniques and acceptable overall technical performance within the operating room, prior to independent practice (Birkmeyer et al., 2013; "CanMEDS 2005 Framework," 2005; Cogbill, 2014; "Intercollegiate Surgical Curriculum Overview," 2013; "Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint," 2012).

The importance of evaluating technical performance within the competency-based education paradigm is essential to ensuring trainees progress through training at the pace that best suits their abilities and needs, however, assessments focusing on technical performance are not well done (Szasz et al., 2015). Furthermore, the current assessment practices utilized internationally are not well documented within the literature, and consequently remain unfamiliar to other stakeholders attempting to implement similar initiatives.

In an era where information dissemination is more feasible than ever, a collaborative effort should allow for the sharing of best practice protocols, in order to further surgical assessment during technical performance. The purpose of this study was to assemble an international EDs' perspective on (1) current technical performance assessment practices (2) ideal technical performance assessment methods and (3) barriers to the adoption of these assessments, at three training stages: selection into training, in-training progression, and certification.

3.3 Methods

An online questionnaire was distributed to EDs internationally using Survey Monkey (Palo Alto, CA). Each question was either formatted as an open-ended response or on a Likert scale from 1-5 (1-strongly disagree, 2-disagree, 3-neutral, 4-agree and 5-strongly agree). For each question the EDs had an opportunity to comment and/or clarify their answers. Each jurisdiction's responses were weighted equally for each question in the survey therefore contributing to $1/8^{th}$ of the results.

EDs (or their equivalent) were deemed most appropriate to participate in the study. EDs are surgeons with major leadership roles in their jurisdictional certifying colleges or official surgical recognition bodies.

Therefore, all of the EDs hold positions of knowledge and authority, having oversight for the certification/examination process for their jurisdiction and the understanding that processes do differ for surgical training, assessment and board recognition internationally. The EDs are all members of the Research, Education and Innovation for Better Outcomes (RETREAT) group, an international consortium for the improvement of surgical training. They are surgeons from

diverse fields with expertise in surgical education and they are responsible for all surgical specialties within their jurisdiction during all stages of training.

Selection

EDs were asked to outline the components of the current surgical selection process in their jurisdictions and their opinions as to whether it is important to test technical aptitude at the selection process. Additional questions were asked to understand how and when the assessment of technical aptitude or skill is being used and furthermore, what type of technical assessment(s) would be appropriate for in-coming trainees in their respective jurisdictions. For the purpose of this study, aptitude was defined as "a natural capacity or ability " to do something ("The Oxford English Dictionary Online, Oxford University Press," March 2015).

In-training and certification

The questions for in-training and certification focused on the assessment of technical competence. At these two time points the study solicited the opinions of EDs as to whether it is important to assess the technical competence of surgical trainees and if there are current assessment practices in place. Moreover, ideal technical performance assessment methods were also sought, (i.e. 'when' and 'where' these assessment should be completed). Finally, EDs opinions were sought on which assessment methods were most appropriate to determine technical competence. For the purpose of this study, competence was defined as "sufficiency of qualification; capacity to deal adequately with a subject" ("The Oxford English Dictionary Online, Oxford University Press," March 2015).

Barriers to technical assessment

Barriers to the implementation of technical performance assessments at all three training stages were also investigated.

3.4 Results

Eight responses were received from a possible ten EDs, with representation from Canada, the United Kingdom, Ireland, Denmark, Hong Kong, Sweden, the Netherlands and Australia & New Zealand.

Selection

Currently selection processes utilize a curriculum vitae, portfolio or application, references, interviews, internship performance scores and technical aptitude. EDs expressed a divided response as to whether technical performance should be assessed during selection with 50% neutral and 50% stating either agree or strongly agree. Similar responses were reported even if an objective measure to assess technical ability prior to entry into training were available (50% neutral and 50% agreeing or strongly agreeing). However, if a technical test were to be incorporated into the selection process the majority of EDs agree that the task should be tailored to basic or intermediate, rather than advanced difficulty.

Of the eight countries, only the United Kingdom reported that the majority of their surgical programs (8 of 10) assess technical ability prior to entry into surgical training and Australia & New Zealand and Ireland report that some programs assess technical ability. In these countries the goal is to assess technical potential rather than a previous learned technical skill and therefore ability is evaluated with either a surrogate test or simulated technical task. The surrogate test is a paper or computer test where a score is generated, whereas, the simulated task is scored under direct preceptor observation using a global rating scale. Either assessment method is reported to contribute as a binary outcome (completed yes or no) or up to 15% of the overall weighted applicant score. The other remaining jurisdictions do not assess technical ability at the time of selection.

In-training

Among EDs 100% either agree or strongly agree, that it is important to objectively assess and document the technical competence of residents during training. Three EDs report that all surgical specialties within their jurisdiction assess technical competence, four responded that some assess technical competence, whereas a single ED reported that none of their surgical specialties assess technical competence during surgical training. Of the seven countries that assess technical competence, the majority (5) use observational assessments and do so in both the operating room and simulation laboratory. Additional methods the EDs felt could ideally be used for technical assessment are represented in Figure 5. Preferred assessment locations include: the operating room (7) over the surgical skills lab (5). During training the majority of

EDs agree or strongly agree that an evaluation for technical competence should be completed at the midpoint of each rotation (6) and at the end of each rotation (8). Whereas, technical competence evaluations at the end of each year for the purpose of promotion were less appealing (5).

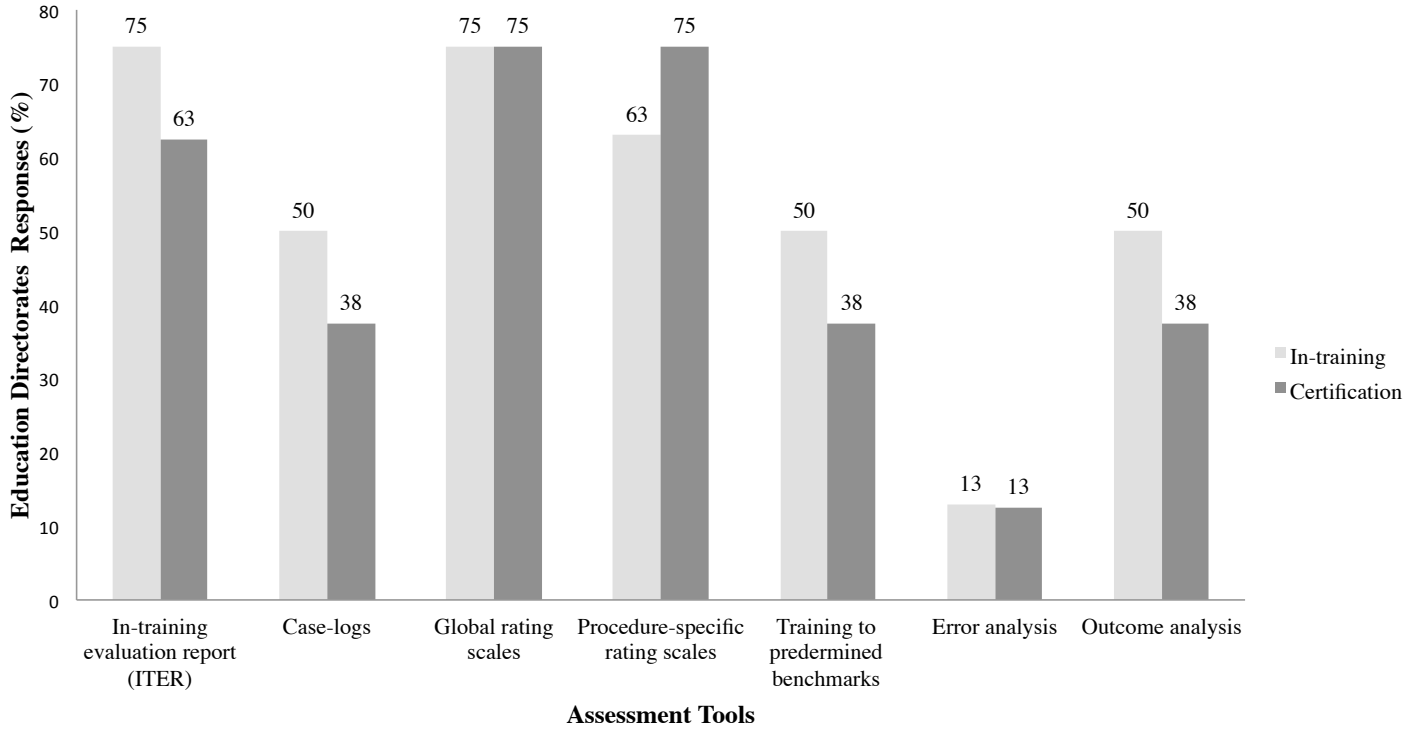

Figure 5. Methods for the assessment of technical competence during surgical training and at certification.

Certification

Among EDs 75% either agree or strongly agree and 25% were neutral as to whether it is important to objectively assess and document the technical competence of residents as part of the certification process. All EDs report that technical competence is currently not objectively assessed as part of their certifications process as a separate examination (8); however, many countries report having to sign off on technical competence prior to certification. The methods

the EDs felt could ideally be used for technical assessment are represented in Figure 5. Preferred assessment locations include: the operating room (6) over the surgical skills lab (3).

Barriers to technical assessment

The major barriers to the adoption of assessments differ depending on the time point of training. Common barriers include: lack of evidence for use in selection (5), financial limitations for test administration during training (4) and a combination of both with respect to certification (3 each) (Figure 6).

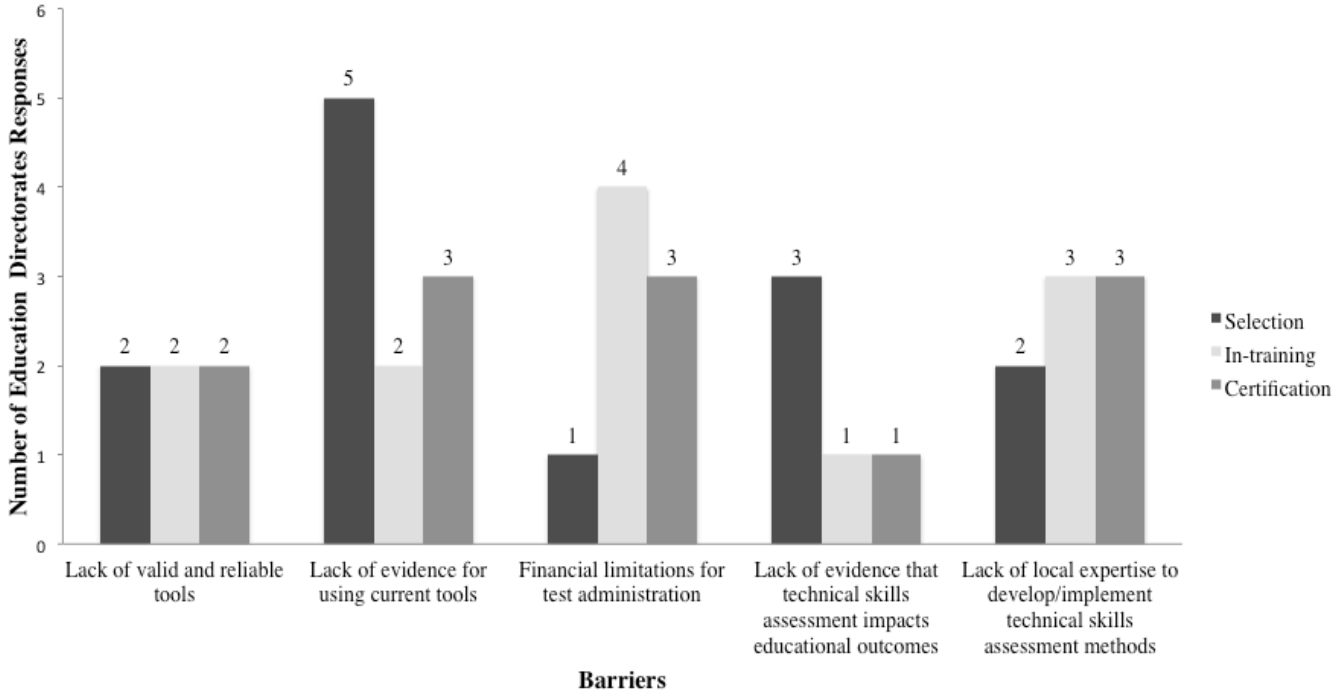

Figure 6. Barriers to incorporating technical assessments during selection, in-training and certification (EDs selected all options that applied to each time-point).

3.5 Discussion

Using a questionnaire, this study explored the current and ideal assessment methods for technical performance along the continuum of surgical training internationally. Furthermore, real or perceived barriers to the administration of such assessments were also determined. This information is crucial as a starting point to share best practice initiatives and further technical performance assessments as the international surgical community shifts training towards a

competency-based education model (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Iobst et al., 2010; "Milestones," 2014; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014).

Technical aptitude assessments during selection are met with resistance due to the limited evidence supporting their predictive utility. The literature suggests that 5-15% of surgical trainees will have difficulty attaining technical competence during residency (Cushchieri, 2003; Grantcharov & Funch-Jensen, 2009). Therefore, identifying these trainees before they embark on a costly (Babineau et al., 2004), rigorous and high stress surgical career is a worthwhile endeavor (Arora, Diwan, & Harris, 2013). Surrogate tests, designed to assess innate abilities that are essential for technical performance, have been studied extensively as screening tools. These tests require few resources to administer and are easily evaluated generating a score that can be directly compared between applicants (Ekstrom, French, Harman, & Dermen, 1976). For these reasons, tremendous effort has been dedicated to searching for the link between surrogate tests such as visual spatial, dexterity and previous experience outside the surgical environment and technical performance. However, increasing evidence suggests that these surrogate tests, do not reliably predict overall future technical performance in the simulation setting or real operating room (M. Louridas, Szasz, de Montbrun, Harris, & Grantcharov, 2016).

At best, surrogate tests may predict a specific subset of skills within a surgical specialty rather than compatibility to surgery in general. PicSOR (Pictorial Surface Orientation), a 2D-3D visual spatial test designed to test laparoscopic potential has been shown to predict technical performance in some laparoscopic tasks but not for endoscopic tasks (Gallagher, Cowie, Crothers, Jordan-Black, & Satava, 2003; Ritter, McClusky, Gallagher, Enochsson, & Smith, 2006). In contrast, the mental rotation test has demonstrated predictive utility in open tasks within plastic surgery only, despite both these surrogate markers designed to test visual spatial ability (Wanzel, Hamstra, Anastakis, Matsumoto, & Cusimano, 2002). Therefore, their application may be narrower than initially expected and caution should be taken not to over generalize their utility in predicting surgical potential.

An alternate approach may be to utilize technical simulated tasks during selection rather than surrogate tests. However, unlike surrogate tests these tasks require costly equipment, and increased resources to generate meaningful performance assessments. To date, direct

assessments by experienced personnel using global rating scales have been used. The department of Otolaryngology at Mayo Clinic (Rochester, MN) has candidates complete a simulated suturing task under the microscope as part of their selection process, which is rated by a faculty member using a global rating scale (Carlson, Archibald, Sorom, & Moore, 2010). This test has been reported to predict future technical performance in their trainees (E. J. Moore, Price, Van Abel, & Carlson, 2014). In the described testing environment, these assessments have not been blinded and therefore are subjective in nature thus at risk of bias. However given that video ratings are even more resource heavy, live assessments are likely a more feasible alternative. Before incorporating simulation tests into selection, further studies are required to identify appropriate tasks for different surgical disciplines followed by a longitudinal follow-up to ensure these tasks predict performance within the clinical environment. Nonetheless, although limited, the initial literature is promising and given the tremendous advances in simulation and robust evidence to support its effectiveness for the transfer of skills into the operating room (Palter et al., 2013), incorporating simulation into selection to predict technical performance may be a logical next step in expanding the benefit of simulation technology.

Assessing the technical competence of trainees during training was seen as important by all of the EDs, more so than during selection or at the time of certification. This is in keeping with recent literature, that states performance assessments should occur at a stage of training where any identified issues can be addressed and remedied, but not so early where the evidence of their utility are lacking or so late that there is no time to appropriately intervene (M. Louridas et al., 2016; Wu, Siewert, & Boiselle, 2010).

Currently, the majority of jurisdictions assess the technical competence of their trainees using subjective methods and/or surrogate markers. Faculty observation of trainee performance has previously come under scrutiny (Elliot & Hickam, 1987; Kalet, Earp, & Kowlowitz, 1992). Multiple studies have demonstrated deficiencies in faculty assessments across various settings, including the OSCE (Kalet et al., 1992). This is often further complicated in the surgical setting as the staff surgeons completing the ratings often work directly with the trainees over prolonged periods of time, leading to subjective ratings influenced by many other factors (Wanzel, Ward, et al., 2002). Surrogate markers have also long been used to infer technical performance (Szasz et al., 2015). Multiple such markers exist and include assessing trainee procedure case-logs, ITERs and training to pre-defined benchmarks (e.g. the Fundamentals of Laparoscopic Surgery (FLS)

examination). Each of which have inherent shortcomings as methods to determine technical competence. Using procedure case-logs can give the number of exposures for a particular case, but since it relies on self-reported data, it is inaccurate and subject to false reporting, furthermore, it is not a direct assessment of technical competence (Lonergan, Mulsow, Tanner, Traynor, & Tierney, 2010). The ITER has been shown as an imprecise method for determining a trainee's technical competence when compared to objective assessments for the same trainees, especially those that are borderline or inadequate, while the FLS examination is a standardized assessment of particular laparoscopic skills and not an assessment of technical competence (L. S. Feldman et al., 2004; Peters et al., 2004). The most appropriate way to circumvent both of these issues is to move away from inferring technical competence towards directly observing trainee performance in the work-place over multiple occasions, completed by well trained and objective evaluators (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Szasz et al., 2015).

The EDs provided a mixed response as to where technical performance assessments should take place, with the operating room slightly outnumbering the surgical skills laboratory. Although the literature on the assessment context has also been mixed, more recent studies have documented that not only are assessments completed in a real world setting (i.e. operating room) important in competency-based education, but they also provide a better platform for trainee feedback and learning (C. Carraccio et al., 2002; Holmboe et al., 2010; Williams et al., 2003). Although simulation will undoubtedly continue to have a large role in competency-based education as suggested by Holmboe *et al*., assessments of competence should occur within a genuine work place setting, or perhaps in combination using both a simulated and real world setting (Holmboe et al., 2010). Moreover, the EDs felt that technical competence assessments should take place at the mid and end-points of each rotation compared to the end of each year for promotion. Although quite appealing from a financial and resource setting, to have low-stakes (formative) assessments occur during rotations for feedback and learning, high-stakes (summative) assessments within the competency-based curricula are also needed, to make decisions about trainees (i.e. trainee progression) (Holmboe et al., 2010; "Intercollegiate Surgical Curriculum Overview," 2013). These high-stakes assessments should utilize standard setting methodologies to ensure the results (decisions) are credible and defensible (Holmboe et al., 2010; J. J. Norcini, 2003).

Although there is interest in including a technical performance examination at the time of certification, this area of assessment is new to the surgical community. Currently, only two examinations have been developed for the purpose of surgical certification (S. L. de Montbrun et al., 2013; van Bockel et al., 2008). The European Board of Vascular Surgery and the American Society of Colon and Rectal Surgeons have developed high-stakes technical performance examinations that take place in the surgical skills laboratory (S. L. de Montbrun et al., 2013; van Bockel et al., 2008). Currently both these certification exams use a version of the objective structure assessment of technical skills (OSATS) global rating scale to assess the candidates. This is inline with EDs responses, where global rating scales and surgical checklists were identified as ideal methods of assessment. However, studies report that global rating scales are superior to checklists when used for summative assessments therefore in the setting of certification; global rating scales may be more appropriate (Regehr, MacRae, Reznick, & Szalay, 1998). These global rating scales are inherently somewhat subjective because personal judgment is required by the examiner to rank the candidate. To counteract this subjectivity, experts from different institutions are recruited to assess the trainees for their final examination; ensuring they do not have an established working relationship with the trainee and thus a predetermined assessment of their technical performance. Both the Vascular and Colon and Rectal Surgery certification examinations are completed in the simulated setting. EDs identified the operating room as the preferred assessment location; however, simulation does ensure that each candidate has the opportunity to be tested on the same set of technical performance markers, with the same scoring system in a standardized controlled environment, which is not the case in the operating room.

One of the major gaps in knowledge is whether passing these technical performance examinations predicts future technical competence. Furthermore, understanding how performance in these examinations translates into real life patient care is also imperative but presently lacking. Therefore, longitudinal studies with the established examinations would be of great benefit to understanding the usefulness of technical assessment at the time of certification.

Barriers to implementing technical performance assessments varied, based on the time-point of training. For the in-training time-point, financial barriers were seen as most common. One way to circumvent such financial barriers is to demonstrate that improved technical performance by residents (as a product of their assessment) ultimately leads to improved patient

outcomes and financial savings (i.e. shorter patient hospitalizations and fewer re-admissions). Such research demonstrating improved patient outcomes with improved performance has already been completed for Staff Surgeons (Birkmeyer et al., 2013). At the selection time-point, increasing the evidence for the utility of such assessments, and in particular demonstrating that the assessment has predictive validity is required prior to implementation (M. Louridas et al., 2016). Finally, for the certification time-point not only is it important to address the financial barriers, but also the resource (expertise) barriers. In particular, it is important to gather keen, adept and trained assessors in order for the assessments to be feasible and the results valid and reliable; properties that are undoubtedly important for all three time-points (Russ et al., 2012).

We acknowledge that the magnitude of the impact of the identified barriers reported in the study may vary between training program within North America and internationally. The variable impact of cost and objective assessment in each program were not explored in this study. Therefore, before potential solutions are sought for programs to adopt similar strategies for technical skills assessment, further research is required to clarify the effect of cost on each surgical program and each jurisdiction as well as the feasibility of introducing objective assessments for technical performance.

3.6 Conclusion

Technical performance assessments are required for the implementation of competency-based education. Technical assessment processes during selection and certification are faced with the uncertainty as to whether these evaluations are predictive of future performance. Future research in these areas should place an added emphasis on longitudinal follow-up to ensure these assessments play a role in the clinical environment whether during training or independent practice. Technical assessment during training has received the most attention, however, these methods are generally subjective in nature and are therefore prone to bias. Global rating scales are likely best suited for the assessment of technical competence during training and moving towards blinded objective ratings, where possible, will further improve these evaluations. Next steps include further research into standard setting methodologies especially for summative assessments. This international awareness as to the current status and future directions of technical assessment will allow training programs to implement best practice assessment methods, while international experts in education research can strive towards developing the

evidence to fill the identified gaps and improve the outlined existing assessment methods at all three training points.

CHAPTER 4

CONSENSUS-BASED TRAINING AND ASSESSMENT MODEL FOR GENERAL SURGERY

Chapter purpose

This chapter provides a consensus-based training and assessment model specific to junior and senior level trainees for operative procedures and tasks in General Surgery. This information can be utilized in General Surgery, as the transition to CBME and assessment occurs to address and remedy the limitations inherent in the current education paradigms. This chapter addresses the limitations identified in section 1.2.8 of Chapter 1.

Chapter preface

The contents of this chapter have been modified from the original version published in the *British Journal of Surgery* and reprinted with permission from BJS Society Ltd Published by John Wiley & Sons Ltd © as: Consensus-Based Training and Assessment Model for General Surgery. **P. Szasz**, M. Louridas, S. de Montbrun, K.A. Harris, T.P. Grantcharov (2016) *BJS* (E pub ahead of print) DOI: 10.1002/bjs.10103

4 Consensus-Based Training and Assessment Model for General Surgery

4.1 Abstract

Background: Surgical education is transitioning into a competency-based education paradigm with the implementation of in-training milestones. At this critical education juncture, it is time to re-evaluate the current General Surgery training guidelines by aligning them with these changes and determining specific procedures for such milestone assessments. The objectives of this study are to outline a consensus in General Surgery on (1) the operative procedures and tasks that are appropriate for junior and senior level trainees, and (2) the procedures that can be used as technical milestone assessments for trainee progression.

Methods: A Delphi technique was distributed to all 17 Canadian General Surgery program directors. Items were ranked on a 5-point Likert scale with consensus defined as Cronbach's $\alpha \geq 0.70$. Items rated ≥ 4 on the 5-point Likert scale by 80% of the program directors were included in the models.

Results: Two Delphi rounds were completed, with 14 program directors taking part in round one and 11 in round two. The overall consensus was high (Cronbach's $\alpha = 0.98$). The training model includes 101 unique procedures and tasks; 24 specific to junior trainees, 68 specific to senior trainees and nine appropriate to all trainees. The assessment model includes four procedures.

Conclusion: We elucidated a training model for operative procedures and tasks for junior and senior level trainees, also describing those not included, information that can serve as a starting-point in creating new guidelines. Moreover, we determined an assessment model for trainee progression, which can be utilized for competency-based assessments in the form of milestones.

4.2 Introduction

Surgical education is changing as a result of a number of complex and interdependent factors, including trainee work hour restrictions, decreased operating room accessibility, increased patient complexity, the growing use of technology, and the recent focus on surgeon volume-outcome data (Barden et al., 2002; Drolet et al., 2013; Jeong et al., 2014; Pofahl & Pories, 2003; Ricci et al., 2014; Rigberg et al., 2000; Traynor, 2011).

As a result, education bodies, including the RCPSC, the ISCP and the ACGME, are implementing competency-based education frameworks with a focus on milestones (J. R. J. Frank, M. et al. , 2005; "Intercollegiate Surgical Curriculum Overview," 2013; "Milestones," 2014). At this critical education juncture, it seems an appropriate time to re-evaluate the current General Surgery training guidelines that underpin residency programs and to begin discussing how to implement more progressive guidelines, which take such recent changes into consideration to ensure we continue to appropriately educate future surgical trainees (M. F. Brennan & Debas, 2004; "Curriculum outline for General Surgery Residency," 2013-2014; "The Future of General Surgery: Evolving to meet a changing practice," 2014; "Objectives of Training in the Specialty of General Surgery," 2010; Pellegrini, 2006; "Specialty of General Surgery Defined," 2013). Furthermore, it is time to seek out procedures that can be used for these technical milestone assessments within the competency-based education framework. As surgical training evolves, the goal of training programs should be to optimize the available operative time with procedures and tasks trainees will perform on a regular basis in order to increase their competence and confidence (Coleman, Esposito, Rozycki, & Feliciano, 2013; Fonseca, Reddy, Longo, & Gusberg, 2014; "The Future of General Surgery: Evolving to meet a changing practice," 2014). While training should be directed away from focusing (often inadequately) on procedures and tasks trainees will not perform after the completion of a General Surgery residency; as these complex surgical procedures are being directed to higher volume centers in order to achieve improved patient outcomes, where fellowship training is all but required (Birkmeyer et al., 2013; Mattar et al., 2013).

The objectives of this study are to utilize a consensus methodology to outline a training model for General Surgery on (1) the operative procedures and tasks that are appropriate for junior and senior level trainees and (2) an assessment model on the procedures that can be used

for technical milestone evaluations, in order to determine whether junior level trainees can progress to senior level trainees.

4.3 Methods

Ethics

The research ethics board (REB) at the University of Toronto approved this study.

Delphi technique

The Delphi technique is a commonly employed methodology for attaining consensus among a group of content experts while minimizing interpersonal interactions in areas where there is a lack of evidence (Fink, Kosecoff, Chassin, & Brook, 1984; Goodman, 1987; Hsu, 2007; J. Jones & Hunter, 1995; Marchant, 1988; Nathens et al., 2003). The major cornerstones of the Delphi technique are fourfold and include: (1) the use of experts, (2) participant anonymity, (3) iterations with interim feedback, and (4) statistical aggregation of results (Goodman, 1987; Hsu, 2007; Rowe, 1999). Although the classic Delphi technique has been used for over 50-years, multiple variations have been developed, each of which has modified one or more of the cornerstones described above (Martino, 1993; Rowe, 1999). There is agreement however, as described by Cuhls *et al*., that the Delphi technique is a survey administered to experts in two or more iterations, in which after the first round the answers (or results) are returned to the experts as a method of interim feedback (Cuhls, 2005).

Participating experts

Experts in the context of a Delphi technique are individuals who understand a topic of study more than others within the same field (Martino, 1993). The experts employed in this study were the General Surgery program directors across Canada within an accredited RCPSC postgraduate training program. Program directors were chosen as the panellists in this Delphi technique as they are experts in training General Surgery residents, they are up to date on changes undertaking their specialty and they are aware of the evolving educational initiatives which have placed an emphasis on achieving specific competencies and milestones (J. R. J. Frank, M. et al. , 2005).

Anonymity

In the Delphi technique, anonymity refers to the participants being unaware of which experts provided which answers (Martino, 1993). Anonymity was maintained in our study as all interactions between the participants took place via an online platform. The benefits of anonymity include a lack of an overly influential participant and the ability of participants to change their opinions without the other participants' knowledge (Goodman, 1987; Martino, 1993; Rowe, 1999).

Multiple iterations with interim feedback

Two rounds of the Delphi technique were completed in order to achieve consensus. The first round content was generated by the program directors then supplemented with a review of the literature and current guidelines ("Curriculum outline for General Surgery Residency," 2013-2014; "Objectives of Training in the Specialty of General Surgery," 2010; "Specialty of General Surgery Defined," 2013). All the responses and associated comments were re-distributed to the program directors for ranking in the second round.

Statistical aggregation of results

The Delphi technique combines the responses of all participating experts to give an overall opinion of the entire group (Martino, 1993). This is accomplished using statistics, in the form of a median and interquartile range (IQR), for each response item (Goodman, 1987; Martino, 1993; Rowe, 1999).

Administration of the Delphi

The Delphi survey was administered using the online platform SurveyMonkey™ (Palo Alto, CA, USA). All program directors were sent an e-mail to their secure university affiliated addresses with instructions and a link to participate. Participation was voluntary. The first round closed six weeks after our initial contact and after three e-mail reminders had been distributed. The second round was returned to only those program directors who responded in the first round. Procedures and tasks suggested to be out of the scope of practice of General Surgery trainees by both a lack of first round responses and current guidelines (i.e. Vascular Surgery, Thoracic Surgery, Transplantation etc.) ("Curriculum outline for General Surgery Residency," 2013-2014;

"Objectives of Training in the Specialty of General Surgery," 2010; "Specialty of General Surgery Defined," 2013) were not included in the second round. The second round closed six weeks after it was distributed and after a subsequent three e-mail reminders.

The first objective of this study was to outline a consensus on the procedures and tasks that are appropriate for junior and senior level trainees in General Surgery. This was accomplished by asking the program directors which procedures and tasks, characterized by anatomic region categories (i.e. upper gastrointestinal tract, lower gastrointestinal tract, breast etc.), they feel a junior trainee (end of PGY-2) and a senior trainee (end of PGY-5) should be competent to perform. Competence in the context of this study was defined for the program directors as "sufficiency of qualification" by the Oxford English Dictionary 2nd edition ("The Oxford English Dictionary," 1989); the ability to safely and efficiently perform that task or procedure without guidance (Szasz et al., 2015).

The second objective of this study was to outline a consensus on the procedures that can be used for technical milestone assessments, in order to determine whether junior level trainees can progress to senior level trainees. This was accomplished by asking the program directors to identify the procedures (performed by all levels of trainees) where variations in performance are most often exhibited (especially as trainees progress through training).

Ranking and consensus determination

All question items were ranked on a 5-point Likert scale. Consensus was reached when Cronbach's α was ≥ 0.70 per question. Consensus guidelines for the Delphi methodology range within the literature (Bland & Altman, 1997; Shrout & Fleiss, 1979) however, the most widely accepted guidelines suggest an internal consistency (measured using Cronbach's α), of ≥ 0.70 for research/education recommendations (Bland & Altman, 1997; Graham, Regehr, & Wright, 2003; Shrout & Fleiss, 1979). In essence, the closer Cronbach's α is to 1.0, the higher the consistency in responses by the expert participants, signifying consensus (Graham et al., 2003).

Inclusion in final models

After consensus was achieved for each question (Cronbach's $\alpha \geq 0.70$), the subsequent step was to determine which items to include in the final models by assessing for positive,

negative or neutral consensus. Items ranked on the 5-point Likert scale as 5 (Strongly agree) or 4 (Agree) by ≥ 80% of the program directors reached positive consensus and were included in the final models (Palter, MacRae, & Grantcharov, 2011; Zevin, Levy, Satava, & Grantcharov, 2012). The procedures and tasks that ranked 1 (Unimportant) or 2 (Less important) by ≥ 80% of the program directors reached negative consensus and were not included in the final models. Finally, all other combinations of rankings received neutral consensus and again, were not included in the final models. Neutral consensus does not signify that every expert member rated the procedure or task as neutral for inclusion or exclusion. Rather, that the group consensus as a whole, did not feel strongly to include or exclude that procedure or task in the final model.

Statistical analysis

Internal consistency for all data collected was determined using Cronbach's α (Bland & Altman, 1997). For each Delphi survey item, the median and IQR of the program directors' responses was determined. SPSS version 22.0 (IBM SPSS Statistics, IBM Corp., Armonk, NY, USA) was used for statistical analysis.

4.4 Results

Participants

Fourteen program directors took part in round one and 11 program directors took part in round two. The participating programs and program director's specialization are shown in Table 4.

Table 4. Participating programs and program director's specialization

1. **University of British Columbia** - General Surgery, Minimally Invasive Surgery (MIS), Surgical Oncology
2. **University of Alberta** - General Surgery, Trauma/Acute Care Surgery
3. **University of Calgary** - General Surgery, Surgical Oncology, Trauma/Acute Care Surgery
4. **University of Saskatchewan** - General Surgery, Critical Care Medicine (CCM)
5. **University of Manitoba*** - General Surgery, Trauma/Acute Care Surgery, Critical Care Medicine (CCM)
6. **Western University** - General Surgery, Colon and Rectal Surgery, Trauma/Acute Care Surgery
7. **University of Toronto** - General Surgery, Trauma/Acute Care Surgery, Critical Care Medicine (CCM)
8. **McMaster University*** - General Surgery, Hepatopancreaticobiliary Surgery
9. **Queens University** - General Surgery, Hepatopancreaticobiliary Surgery, Minimally Invasive Surgery (MIS)
10. **University of Ottawa** - General Surgery, Colon and Rectal Surgery
11. **University of Montreal*** - General Surgery, Surgical Oncology
12. **Sherbrooke University** - General Surgery, Minimally Invasive Surgery (MIS)
13. **Dalhousie University** - General Surgery, Hepatopancreaticobiliary Surgery, Transplant, Trauma/Acute Care Surgery
14. **Memorial University** - General surgery, Minimally invasive surgery (MIS)

* These programs completed part one of the Delphi t only.

A Consensus training model for operative procedures and tasks

The final consensus training model includes 101 unique procedures and tasks, both open and minimally invasive. Twenty-four procedures and tasks are specific to junior level trainees (Table 5), 68 are specific to senior level trainees (Table 6) and nine are common across all trainees (identified by a * in Table 2 and Table 3). These procedures and tasks can be categorized by anatomic region and include: upper gastrointestinal (3), lower gastrointestinal (24), hepatopancreaticobiliary (13), hernia (16), perianal (13), breast (11), soft tissue (5), and trauma and emergency (16) representation. Overall, consensus was found to be high with a Cronbach's $\alpha = 0.98$. Individual question consensus ranged from Cronbach's α's 0.69 to 0.95. All questions, except junior trainee '*Soft tissues procedures and tasks'*, reached consensus.

Only a single procedure, the laparoscopic mobilization of the rectum, for junior trainees in the '*Open and minimally invasive lower gastrointestinal tract procedures and tasks*' category demonstrated negative consensus, representing no reason for inclusion in the final model.

Finally, 68 procedures and tasks, both open and minimally invasive, received neutral consensus, signifying that program directors did not strongly feel either way (in a positive or negative direction) that these should be included in or excluded from the final training model (Table 7 and Appendix 2). Thirty of these procedures and tasks are specific to junior trainees and 38 are specific to senior trainees. These procedures and tasks were embedded within each anatomic region category. The '*Open and minimally invasive endocrine procedures and tasks'* for both junior and senior trainees had all of its procedures and tasks fall within the neutral consensus group and included: opening and closing thyroid incision, thyroid fine-needle aspiration (FNA), total thyroidectomy, subtotal thyroidectomy, hemi-thyroidectomy, parathyroidecomy and adrenalectomy (open and laparoscopic).

Table 5. Junior level procedures and tasks included in the final consensus training model, arranged by anatomic category

Anatomical categorization of tasks and procedures	Cronbach's α per question	Median (IQR‡) per item
1. Open and minimally invasive upper gastrointestinal tract	**0.75**	
Insertion of laparoscopic trocars		**5.0 (1.0)**
Use of surgical staplers		**5.0 (1.0)**
2. Open and minimally invasive lower gastrointestinal tract	**0.93**	
Closure of stoma*		**4.0 (0.25)**
Laparoscopic appendectomy - simple*		**5.0 (0.25)**
Open appendectomy - simple*		**5.0 (1.0)**
3. Open and minimally invasive hepatopancreaticobiliary	**0.85**	
Opening/closing chevron incision		**5.0 (1.0)**
Clip application to cystic artery/duct		**4.0 (1.0)**
Laparoscopic mobilization of gallbladder off liver		**4.0 (1.0)**
Laparoscopic cholecystectomy – simple*		**4.0 (1.0)**
4. Open and minimally invasive hernias	**0.95**	
Inguinal hernia incision		**5.0 (1.0)**
Inguinal hernia mobilization of spermatic cord		**5.0 (1.0)**
Inguinal hernia sac identification		**4.5 (1.0)**
Primary repair of inguinal hernia - simple*		**4.0 (1.25)**
Mesh repair of inguinal hernia - simple*		**4.0 (1.0)**
Umbilical hernia incision		**5.0 (1.0)**
Umbilical hernia sac identification		**5.0 (1.0)**
Umbilical hernia repair		**4.5 (1.0)**
Ventral hernia incision		**4.0 (1.25)**
5. Perianal	**0.87**	
I&D anorectal abscess*		**5.0 (1.0)**
EUA*		**5.0 (1.0)**
Fistula identification		**4.0 (1.0)**
Hemorrhoid banding*		**4.0 (0.25)**
Excision pilonidal sinus		**4.0 (0)**
Flexible sigmoidoscopy		**4.0 (1.0)**
Rigid sigmoidoscopy		**4.5 (1.0)**

6. Breast	**0.90**	
I&D breast abscess		**4.5 (1.0)**
Closure of mastectomy incision		**5.0 (1.0)**
Breast biopsy (FNA)		**4.0 (1.0)**
Breast biopsy (surgical)		**4.0 (1.0)**
7. Emergency and trauma	**0.83**	
Opening/closing of laparotomy		**5.0 (1.25)**
FAST		**4.0 (0.5)**
Chest tube placement		**5.0 (1.0)**
Central venous catheter placement		**5.0 (1.0)**
8. Soft tissue	**N/A†**	
I&D subcutaneous abscess		**5.0 (0.25)**
Excision of benign skin and soft tissue lesion		**5.0 (1.0)**

* Tasks and procedures appropriate for both junior and senior trainees
† Level of consensus not reached, not included in final model (please see text for full explanation)
‡ IQR= Q3-Q1 (The difference between the third and first quartiles)
I&D - incision and drainage, EUA - evaluation under anesthesia, FNA - fine needle aspiration,
FAST – focused assessment with sonography

Table 6. Senior level procedures and tasks included in the final consensus training model, arranged by anatomic category

Anatomical categorization of tasks and procedures	Cronbach's α per question	Median (IQR‡) per item
1. Open and minimally invasive upper gastrointestinal tract	**0.92**	
OGD		**5.0 (1.0)**
2. Open and minimally invasive lower gastrointestinal tract	**0.93**	
Laparoscopic mobilization of right colon		**4.5 (1.0)**
Open mobilization of right colon		**5.0 (0.25)**
Laparoscopic mobilization of sigmoid/ left colon		**4.5 (1.0)**
Open mobilization of sigmoid/ left colon		**5.0 (0.25)**
Laparoscopic mobilization of rectum		**4.0 (1.25)**
Open mobilization of rectum		**5.0 (0.25)**
Open splenectomy		**4.5 (1.0)**
Laparoscopic right hemicolectomy		**4.5 (1.0)**
Open right hemicolectomy		**5.0 (0.25)**
Laparoscopic left hemicolectomy		**4.5 (1.0)**
Open left hemicolectomy		**5.0 (0.25)**
Laparoscopic sigmoid resection		**4.5 (1.0)**
Open sigmoid resection		**5.0 (0.25)**
Open LAR		**4.0 (1.0)**
Laparoscopic small bowel resection		**4.5 (1.0)**
Open small bowel resection		**5.0 (1.0)**
Intracorporeal anastomosis		**4.0 (1.0)**
Intracorporeal vessel ligation (i.e. ileocolic/ IMA)		**5.0 (1.0)**
Creation of loop/end stoma		**5.0 (0.25)**
Closure of loop stoma*		**5.0 (0.25)**
Laparoscopic appendectomy – simple*		**5.0 (0.25)**
Laparoscopic appendectomy – complex		**5.0 (0.25)**
Open appendectomy – simple*		**5.0 (0.25)**
Open appendectomy – complex		**5.0 (0.25)**
3. Open and minimally invasive hepatopancreaticobiliary	**0.82**	
Control of portal structures for bleeding		**4.0 (1.0)**
Kocher manoeuvre		**5.0 (0.25)**
Mobilization of liver		**4.5 (1.0)**

Exposure of pancreas		**5.0 (1.0)**
Laparoscopic cholecystectomy – simple*		**5.0 (0.25)**
Laparoscopic cholecystectomy - complex		**5.0 (0.25)**
Open cholecystectomy - simple		**5.0 (1.0)**
Open cholecystectomy - complex		**5.0 (1.0)**
Open CBD exploration		**4.0 (1.5)**
Open repair of liver injuries - simple		**4.0 (1.25)**
4. Open and minimally invasive hernias	**0.78**	
Primary repair of inguinal hernia*		**5.0 (0.25)**
Primary repair of inguinal hernia - recurrent		**5.0 (0.25)**
Mesh repair of inguinal hernia - simple*		**5.0 (0.25)**
Mesh repair of inguinal hernia - recurrent		**5.0 (0.25)**
Femoral hernia repair		**5.0 (1.0)**
Laparoscopic ventral hernia repair		**4.0 (0.5)**
Open ventral hernia repair		**5.0 (1.0)**
Parastomal hernia repair		**5.0 (1.0)**
Open intra-abdominal hernia repair (i.e. Petersen's space)		**4.0 (0.5)**
5. Perianal	**0.89**	
EUA*		**5.0 (1.0)**
I&D anorectal abscess*		**5.0 (0.25)**
Fistulotomy		**5.0 (1.0)**
Fistulectomy		**5.0 (1.25)**
Seton placement		**5.0 (0.25)**
Hemorrhoidectomy		**5.0 (1.0)**
Hemorrhoid banding*		**5.0 (0.25)**
Lateral internal sphincterotomy		**4.5 (1.0)**
Colonoscopy		**5.0 (0.25)**
6. Breast	**0.78**	
Lumpectomy		**5.0 (0.25)**
Excision of ducts		**5.0 (0.25)**
Mastectomy - simple		**5.0 (0.25)**
Mastectomy- skin sparing		**5.0 (1.0)**
Mastectomy - modified radical		**5.0 (1.0)**
SLNB		**5.0 (1.0)**
ALND		**5.0 (0.25)**

7. Emergency and trauma	**0.91**	
Damage control laparotomy		**5.0 (0.25)**
Abdominal packing		**5.0 (0.25)**
Trauma splenectomy		**5.0 (1.0)**
Surgical management of SBO		**5.0 (0.25)**
Surgical management of LBO		**5.0 (0.25)**
Oversewing of bleeding duodenal ulcer		**5.0 (1.0)**
Laparoscopic repair of duodenal ulcer		**4.0 (1.0)**
Open repair of duodenal ulcer		**5.0 (1.0)**
Laparoscopic repair of enterotomy		**4.0 (0.25)**
Open repair of enterotomy		**5.0 (0.25)**
ED thoracotomy		**4.0 (1.25)**
Surgical airway		**5.0 (1.0)**
8. Soft tissue	**0.84**	
Debridement for necrotizing fasciitis		**5.0 (0)**
Debridement of wounds		**5.0 (1.0)**
Lymph node biopsy		**5.0 (0.25)**
Excision of benign skin and soft tissue lesion		**5.0 (1.0)**
Excision of malignant skin and soft tissue lesion		**5.0 (1.0)**

* Tasks and procedures appropriate for both junior and senior trainees

‡ IQR= Q3-Q1 (The difference between the third and first quartiles)

OGD - esophagogastroduodenoscopy, LAR – low anterior resection, IMA – inferior mesenteric artery, CBD – common bile duct, EUA - evaluation under anesthesia, I&D - incision and drainage, SLNB - sentinel lymph node biopsy, ALND - axillary lymph node dissection, SBO – small bowel obstruction, LBO – large bowl obstruction, ED – emergency department

Table 7. Procedures and tasks excluded from the final consensus training model, arranged by anatomic category

Anatomical categorization of tasks and procedures	Number of procedures	Median (IQR‡) distribution*
1. Open and minimally invasive upper gastrointestinal tract		
Senior trainees	**7**	**2.5-4.0 (1.0-1.75)**
Junior trainees	**3**	**3.0-4.0 (0-0.75)**
2. Open and minimally invasive lower gastrointestinal tract		
Senior trainees	**4**	**3.5-4.0 (1.0-1.5)**
Junior trainees	**9**	**2.0-4.0 (0.75-2.0)**
3. Open and minimally invasive hepatopancreaticobiliary		
Senior trainees	**11**	**2.0-4.0 (0.75-1.75)**
Junior trainees	**4**	**2.0-4.0 (1.0-1.75)**
4. Open and minimally invasive endocrine		
Senior trainees	**6**	**3.0-4.0 (1.0-2.5)**
Junior trainees	**2**	**4.0 (1.75-2.0)**
5. Open and minimally invasive hernias		
Senior trainees	**3**	**3.0 (0-1.75)**
Junior trainees	**1**	**4.0 (1.0)**
6. Perianal		
Senior trainees	**3**	**2.5-3.0 (1.0-1.75)**
Junior trainees	**3**	**3.0-4.0 (0.75-1.75)**
7. Breast		
Senior trainees	**1**	**4.0 (1.75)**
Junior trainees	**4**	**3.0-4.0 (0.75-1.75)**
8. Emergency and trauma		
Senior trainees	**0**	**–**
Junior trainees	**3**	**4.0 (1.5-2.0)**

9. Soft tissue		
Senior trainees	**3**	**3.0-4.0 (1.0-2.75)**
Junior trainees	**1**	**3.0 (1.0)**

‡ IQR= Q3-Q1 (The difference between the third and first quartiles)
* Distribution of the median and IQR ranges for all of the procedures in that category per trainee group

A Consensus model of procedures for technical milestone assessments

The procedures in the final consensus assessment model for the progression of a junior level trainee to a senior level trainee are depicted in Table 8.

Table 8. Procedures for technical milestone assessments

Procedure	Median (IQR*) per item
1. Laparoscopic cholecystectomy	**4.0 (1.0)**
2. Laparoscopic appendectomy	**4.0 (0)**
3. Inguinal hernia repair	**4.0 (0)**
4. Small bowel resection	**4.0 (0)**
	Overall Cronbach's α = 0.71

* IQR= Q3-Q1 (The difference between the third and first quartiles)

4.5 Discussion

Using a consensus-based methodology, the Delphi technique, we elucidated a training and assessment model for General Surgery trainees. This information is imperative as a starting point to initiate a wider dialogue as we embark on implementing competency-based training frameworks and their associated milestones into residency training.

Bell *et al*. previously ascertained the opinions of United States program directors on procedures surgical residents should be competent to undertake upon the completion of residency and compared those to actual case-logs of General Surgery residents – demonstrating a wide disparity (R. H. Bell, Jr., 2009; R. H. Bell, Jr. et al., 2009). Although an excellent reference for end of training objectives, this framework does not offer the granularity necessary to implement trainee level-specific guidelines, nor the possibility for milestone assessments. For the first time, we present a consensus-based training model for operative procedures and tasks appropriate to junior and senior trainees in General Surgery. Moreover, we outline a consensus-based assessment model on operative procedures that can be used as milestone assessments, to determine if trainees are appropriately progressing through their surgical training and ultimately for in-training promotion.

With decreased work hours (Barden et al., 2002; Drolet et al., 2013) and graded resident autonomy (Halpern & Detsky, 2014), as well as the increased use of operative technology (Mattar et al., 2013; Swanstrom et al., 2012), patient complexity (Pofahl & Pories, 2003; Rigberg et al., 2000), and focus on surgeon volume-outcome data (Jeong et al., 2014; Ricci et al., 2014), residency training programs should strive to optimize the available training time with procedures and tasks that trainees will perform on a regular basis at the completion of training in order to increase their competence and confidence (Coleman et al., 2013; Fonseca et al., 2014; "The Future of General Surgery: Evolving to meet a changing practice," 2014).

To this end, our first objective was to outline a consensus on the operative procedures and tasks that are appropriate for junior and senior trainees in General Surgery. The consensus training model presented here focuses on procedures and tasks at both the junior and senior levels that are attainable and necessary in order to progress through training and graduation, without fellowship training.

The procedures and tasks where positive consensus was achieved were included in the final training model. This included all of the major General Surgical categories, divided by anatomic region. These procedures and tasks are in keeping with previous literature published regarding the most common surgical exposures for General Surgery residents (R. S. Chung, 2005) and in regards to the senior residents, the list compiled by Bell *et al*. (R. H. Bell, Jr. et al., 2009). The procedures and tasks included in the final training model also represent the more common procedures performed by practicing General Surgeons ("The Future of General Surgery: Evolving to meet a changing practice," 2014). Their inclusion signifies the program director's agreement of the skill set required by trainees to enter independent practice without fellowship training. This is further demonstrated by the common procedures and tasks performed by General Surgeons ("The Future of General Surgery: Evolving to meet a changing practice," 2014) having the highest median and smallest IQRs within this study, exemplifying the utmost agreement for inclusion in the final training model. Finally, the procedures with the highest median and smallest IQRs in this study are also in keeping with the list compiled by Brennan *et al*. depicting how previously trained General Surgeons (often in the absence of a fellowship) at the time of their 10-year recertification have self-differentiated their own practice into only a small subset of their original training (M. F. Brennan & Debas, 2004). In our study, procedures in each anatomic category reached consensus except junior trainee *'Soft tissues procedures and tasks'*. This was conceivably a result of a true lack of consensus, but more likely because there were only a few item choices (three) for that question. The lack of item choices has been documented in the literature to deflate the true value of Cronbach's α (Cortina, 1993). Given a lack of consensus for '*Soft tissues procedures and tasks'*, they were not included in the final model.

The only procedure achieving negative consensus was the laparoscopic mobilization of the rectum for junior trainees in the *'Open and minimally invasive lower gastrointestinal tract procedures and tasks'* category. The fact that only a single procedure achieved negative consensus is interesting. This likely reflects the program directors' reluctance to completely remove procedures from this new training model. The reasons for this reluctance are likely complex, but may (in part) be a result of the perception held amongst the program directors that any procedure/task, even one performed rarely, may have elements of transferability to other more commonly performed procedures/tasks. The literature on skill transfer, however, is mixed

(Ahlering, Skarecky, Lee, & Clayman, 2003; Figert et al., 2001; Mattar et al., 2013; A. E. Park, Lee, T.H., 2011; Seymour et al., 2002). The lack of excluded procedures in our study, however, is in keeping with the results from Bell *et al*., who documented a total of eight procedures (from a list of 300), the United States program directors did not feel were appropriate for General Surgery resident training (R. H. Bell, Jr. et al., 2009).

Although all of the major General Surgery categories had procedures and tasks that achieved neutral consensus, a few interesting trends developed. Noticeably, the '*Open and minimally invasive endocrine procedures and tasks'* category for both junior and senior trainees had all of its procedures and tasks achieve neutral consensus. The procedures included in this category were all thyroid or parathyroid related except for the adrenalectomy. Program directors did not feel junior nor senior residents should be competent to perform an open/laparoscopic adrenalectomy, which is perhaps not unreasonable, given trainees' minimal operative experience with this procedure during residency and its technical modification compared to traditional laparoscopic surgery (R. H. Bell, Jr. et al., 2009; Walz et al., 2006). Surprising however, was that program directors did not feel strongly that General Surgery trainees should be competent performing thyroid or parathyroid procedures or tasks. This is in contrast to previous literature supporting both an expectation of and experience with thyroid procedures (R. H. Bell, Jr. et al., 2009; R. S. Chung, 2005). The view presented here by the program directors may be in keeping with recent shifts in volume outcome research in many subspecialties of General Surgery, including thyroid surgery and the increasing number of residents undertaking fellowship training prior to independently completing these procedures (Coleman et al., 2013; Fonseca et al., 2014; Sosa et al., 1998). Most procedures and tasks in the *'Open and minimally invasive upper gastrointestinal tract procedures and tasks'* category and the *'Open and minimally invasive hepatopancreaticobiliary procedures and tasks*' category for senior trainees also achieved neutral consensus. In particular, the program directors did not feel a senior resident should be competent to perform major upper gastrointestinal procedures, such as a gastrectomy (total or partial), an anti-reflux procedure or a pyloromyotomy. Furthermore, the program directors did not feel a senior resident should be competent to perform advanced hepatopancreaticobiliary procedures, including pancreatic resections, pancreatic anastomoses and biliary anastomoses. Although likely intuitive that trainees without subspecialized training in advanced minimally invasive and hepatopancreaticobiliary surgery should not be performing such procedures, given

their complexity and steep learning curves, this has not yet been reflected and implemented into currently utilized General Surgery training curricula ("Curriculum outline for General Surgery Residency," 2013-2014; Jeong et al., 2014; "Objectives of Training in the Specialty of General Surgery," 2010; Ricci et al., 2014; "Specialty of General Surgery Defined," 2013). Perhaps what is most useful about this training model is not the procedures and tasks present in the final model, but rather the procedures and tasks that are absent. Keeping these tasks and procedures in mind for the future of curriculum design to the RCPSC, ISCP, ACGME and the SCORE, may prove useful in streamlining the process ("Curriculum outline for General Surgery Residency," 2013-2014; J. R. J. Frank, M. et al. , 2005; "Intercollegiate Surgical Curriculum Overview," 2013; "Milestones," 2014).

The more recent development of milestones (J. R. Frank, Snell, L.S., Sherbino, J, 2014; "Milestones," 2014; Nasca et al., 2012; Paddick, 2010) can serve as the link between competency-based education paradigms and competency-based assessments. By definition, milestones are markers of expected trainee performance across the many stages of training ("Benefits of CBD for medical educators in each specialty," 2014). These milestones are defined in a clear manner and they allow program directors to identify specific trainees that are ready to move onto the next and more complex phase of training (J. R. Frank, Snell, L.S., Sherbino, J, 2014).

To this end, our second objective was to outline an assessment consensus on the procedures in General Surgery that can be used for milestone assessments, in order to determine whether junior level trainees can progress to senior level trainees. The four procedures meeting consensus included: laparoscopic cholecystectomy, laparoscopic appendectomy, inguinal hernia repair and small bowel resection. These procedures can each then become milestones that together need to be attained in order for surgical trainees to transition and progress from a junior to senior trainee (from a technical competence standpoint), in a holistic manner in coordination with the other RCPSC, ISCP and ACGME competencies (J. R. J. Frank, M. et al. , 2005; "Intercollegiate Surgical Curriculum Overview," 2013; "Milestones," 2014). To ensure that these performance assessments are a reliable representation of a trainee's performance, these procedures musts be assessed on multiple occasions, by multiple assessors (Marriott et al., 2011). However, if, technical performance deficits are identified at these milestones assessments, these deficits may be addressed and remedied early, rather than near the completion of training, which

has been shown less effective (Wu et al., 2010). A method to ensure that the competent/non-competent decisions around these milestones assessments are defensible and credible is to set performance standards using standard-setting methodologies, adopted from the education literature (J. J. Norcini, 2003; J. J. Norcini, Shea, JA, 1997; Schindler et al., 2007; Szasz et al., 2015; T. J. Wilkinson et al., 2001)

There are four primary limitations present within our study. First, we had three non-responders (14 of 17 responders) in the first round, with a subsequent loss of another three program directors in the second round (11 of 17 responders). Although this is within the acceptable non-responder and attrition rates in the literature, the reliability of Delphi techniques has been described to increase with higher participant numbers, although no clear minimum number has been documented (de Villiers, de Villiers, & Kent, 2005; Fink et al., 1984; Hsu, 2007; Sierles, 2003; Zevin et al., 2012). It has been shown however, that both responders and non-responders have similar underlying characteristics and that the opinions of the responders are representative of their colleagues who have not responded (McKee, Priest, Ginzler, & Black, 1991). The second limitation is that although program directors are the most appropriate individuals to make decision regarding training guidelines (content validity), the fact that they predominantly work in academic centers may make this list more appropriate to academic residency training programs compared to community-based programs. However, many academic programs have community satellite programs where their residents train. Therefore program directors are actively involved and therefore aware of the operative needs and experiences of both academic and community centers and their role in resident training. The third limitation is that the study was conducted exclusively with Canadian program directors, therefore transferability to international programs can potentially be seen as a concern. However, with the worldwide transition to competency-based education models, programs will likely become more similar and the sharing of best practice protocols will become imperative, even if only as a conversation starter in the broader international community. The fourth limitation relates to the differences in surgical training programs internationally. Specifically, the differences in the structure and duration of more basic training with varying entry and exit points into more subspecialized training. The models presented here may lend more information to programs similar in structure to Canada. Nonetheless, the findings presented here, if not in full, can in part be adopted by many countries for their applicable timelines. Furthermore, these findings start the

discussion about how training and assessment models may be developed within competency-based training, a framework that many countries are moving towards (J. R. J. Frank, M. et al. , 2005; "Intercollegiate Surgical Curriculum Overview," 2013; "Milestones," 2014).

4.6 Conclusion

Using a consensus-based methodology, the Delphi technique, we elucidated a training model for operative procedures and tasks that is appropriate for junior and senior level trainees in General Surgery. Moreover, we determined an assessment model that can be used to evaluate whether junior level trainees can progress to senior level trainees. The former can be utilized by education stakeholders to inform and develop new training guidelines; while the latter can be used to create in-training milestones to assess resident technical competence and resident progression through training. What is perhaps most informative about the training model in particular, is not the procedures and tasks that are present, but rather those that are absent, information that may be fundamental as surgical education transitions into a competency-based paradigm. Nonetheless, all of these findings play an intricate role in the development and implementation of competency-based assessments, which are currently lacking, but essential as we move through the 21st century of surgical training.

CHAPTER 5

SUBJECTIVE AND OBJECTIVE PERFORMANCE ASSESSMENTS CORRELATE IN THE OPERATING ROOM

Chapter purpose

This study identifies internal raters (staff surgeons) who are intimately tied to the training of residents as reliable assessors for technical and non-technical performance assessments in the operating room in comparison to their external rater (no relationship to residents) counterparts. This chapter addresses the limitations identified in section 1.8.5 of Chapter 1.

Chapter preface

The contents of this chapter have been accepted to the *American Journal of Surgery* for publication as: Strategies for Increasing the Feasibility of Performance Assessments During Competency-Based Education: Subjective and Objective Evaluations Correlate in the Operating Room. **P. Szasz**, M. Louridas, K.A. Harris, T.P. Grantcharov (2016)

5 Subjective and Objective Performance Assessments Correlate

5.1 Abstract

Background: Competency-based education necessitates assessments that determine whether trainees have acquired specific competencies. The evidence on the ability of internal raters (staff surgeons) to provide accurate assessments is mixed. This ability has not been explored directly in the operating room. The objective of this study is to compare the ratings given by internal raters versus an expert external rater (individuals independent to the training process) in terms of the performance scores they attribute to trainees in the operating room.

Methods: Both an internal and an external rater assessed General Surgery residents during a laparoscopic cholecystectomy for their technical and non-technical performance, using the OSATS and OSANTS instruments, respectively.

Results: Fifteen cases were observed. There was a moderately positive correlation ($r_s = 0.618$, $p = 0.014$) for technical performance and a strong positive correlation ($r_s = 0.731$, $p = 0.002$) for non-technical performance. Overall, the internal raters were less stringent for both technical (median score difference 2.0; mean rank 3.33 vs. 8.64, $p = 0.007$) and non-technical (median score difference 1.0; mean rank 3.83 vs. 8.50, $p = 0.01$) performances.

Conclusion: Internal raters were positively correlated to the external rater, but less stringent for both technical and non-technical performance. This information is timely as stakeholders look to implement competency-based assessments, which could utilize internal raters, at least for formative assessments. Next steps include evaluating how internal rater stringency can be aligned with that of external raters and subsequently whether internal raters can be utilized for summative assessments.

5.2 Introduction

Competency-based education, an outcome-focused training paradigm, is significantly altering the way in which surgical residents are being trained; with less of a focus on the duration of training and more of a focus on the acquisition and demonstration of specific competencies ("Accreditation Council for Graduate Medical Education (ACGME) - Next Accreditation System (NAS) - Milestones," 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014). Although relatively new, it has permeated every medical specialty worldwide, and it is changing not only the way surgical residents are being educated, but also assessed ("Accreditation Council for Graduate Medical Education (ACGME) - Next Accreditation System (NAS) - Milestones," 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014; Holmboe et al., 2010; "Intercollegiate Surgical Curriculum Programme: Overarching Assessment Blueprint 2012 ", 2012).

Although formative assessments, which are used regularly for trainee learning, development and feedback, are at the center of competency-based education, periodic assessments that determine whether residents have acquired specific competencies are also required (C. Carraccio et al., 2002; Holmboe et al., 2010). These summative assessments are used to assess trainee learning at specific intervals in order to make decisions about that trainee (i.e. promotion, remediation) (Holmboe et al., 2010). Given that summative assessments can have significant implications for residents, training programs and licensing authorities, they must be developed, organized, and implemented in a thoughtful manner (R.J. Mislevy, 2003; R.J. Mislevy, 2011; Wass, Van der Vleuten, Shatzer, & Jones, 2001). One area of focus in relation to assessment practices, particularly important for summative assessments, has been who the assessors ought to be (Holmboe et al., 2010; Swing et al., 2009).

There is an emerging area of literature that evaluates the utility of staff physician assessments (deemed internal assessments for the remainder of the manuscript) for both technical and non-technical resident performance (L. S. Feldman et al., 2004; Herbers et al., 1989). In particular, the literature is mixed as to the accuracy and reliability of internal assessments, compared to other more 'objective' measures of trainee performance (i.e. standardized assessments, licensing examinations) (Awad, Liscum, Aoki, Awad, & Berger, 2002; Farrell et al., 2010; L. S. Feldman et al., 2004; Goldstein et al., 2014; Reid et al., 2014; Van der Vleuten, Norman, & De Graaff, 1991). Despite comparisons to other such 'objective' measures of

performance, a comparison between internal and external assessors (those individuals that are independent to the training process of the resident being assessed) – while often if only anecdotally presumed superior, is lacking. Potential reasons against using internal assessors for summative type assessments include their knowledge of a trainee and previous trainee encounters, as well as the natural conflict of interest that exists between an internal assessor and trainee (the internal assessor is intimately tied to and responsible for training a resident – therefore they are not only assessing trainee performance but also their ability to teach) (T. J. Wilkinson & Wade, 2007). In contrast, potential reasons for using internal assessors for summative type assessments within competency-based education includes their ability to address and possibly minimize some of the logistical issues inherent to such assessment frameworks (i.e. cost) (Norman et al., 1991; Takahashi, 2011).

A study that directly compares internal and external raters in the operating room has yet to be completed. Therefore, the objective of this study is to compare the ratings given by internal raters versus an external rater in terms of the performance scores they attribute to trainees in the operating room during a laparoscopic cholecystectomy. This will be accomplished by evaluating 1) total score correlations and 2) mean rank differences, between internal and external rater attributed scores for both technical and non-technical performance.

5.3 Methods

Ethics

The research ethics boards at the University of Toronto and St. Michael's Hospital approved this study.

Procedure and assessment instruments

The procedure chosen for this study was the laparoscopic cholecystectomy. Resident technical performance was evaluated using the OSATS (J. A. Martin et al., 1997) global rating instrument and resident non-technical performance was evaluated using the OSANTS (Dedy, Szasz, et al., 2015) global rating instrument (Appendix 3 and 4 respectively). Both of these rating instruments have seven categories, each ranked on a 5-point Likert scale with a maximum score of 35 (Dedy, Szasz, et al., 2015; J. A. Martin et al., 1997). In a traditional sense, the OSATS (J.

A. Martin et al., 1997) instrument has previously demonstrated construct validity, inter-rater reliability, and internal consistency and in a more contemporary sense meets many of the criteria within Messick's conceptual framework of validity (S. M. Downing, 2003; Ghaderi et al., 2015; J. A. Martin et al., 1997; Messick, 1995; Reznick et al., 1997). Similarly, the OSANTS (Dedy, Szasz, et al., 2015) instrument has also previously demonstrated construct validity, concurrent validity, inter-rater reliability, and internal consistency and in a more contemporary sense fits many of the criteria within Messick's conceptual framework of validity (Dedy, Szasz, et al., 2015; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995).

Participants

The resident participants were General Surgery trainees at the University of Toronto, completing a rotation at St. Michael's Hospital.

The internal raters were all board certified General Surgeons at St. Michael's Hospital with experience in minimally invasive surgery. They are deemed internal as they are intimately tied to the training of the residents they are assessing, often having previous knowledge of and experience with these residents. These surgeons had some previous informal experience assessing trainee technical performance and little to no experience assessing trainee non-technical performance. Prior to any assessments taking place, each internal rater underwent formalized training. A combination of the Performance Dimension Training (PDT) and Rater Error Training (RET) strategies were utilized (M. Feldman et al., 2012). During a one-hour tutorial, the specific constructs under study for each rating scale item on the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) rating instruments was explained and discussed, and examples of good and bad markers of performance were described for each (M. Feldman et al., 2012). Furthermore, the internal raters attention was drawn to rating errors for each scale item and again this was discussed (M. Feldman et al., 2012). Finally, after the first live assessment, internal raters underwent a debriefing session with a member of the research team also present in the operating room, where the case was discussed and their OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) ratings explored/compared to that of the research team member for calibration purposes (M. Feldman et al., 2012).

The external rater was a surgical education and assessment expert, previously undergoing formalized training similar to the abovementioned, in both technical and non-technical assessment using the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) global rating instruments, with at least a 100 case experience for each. This rater is deemed external, as he is independent to the training process of the residents. In this study, the external rater was considered to be the 'gold standard' given the extent of training, calibration with other raters, and case experience he had.

Data collection

All data collection occurred in the operating room at St. Michael's Hospital. After the completion of a laparoscopic cholecystectomy by a resident, the internal and external rater separately filled out the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) instrument ratings to assess the technical and non-technical performance.

Sample size

An a priori sample size calculation, with an alpha = 0.05, power = 0.95 and effect size r = 0.72, revealed that 15 case observations were required (G*power samples size software [Institute for Experimental Psychology, Dusseldorf Germany]). The effect size is based on studies comparing untrained and trained raters, where the correlations between trained raters ranged from 0.72 to 0.76 (Sevdalis et al., 2009).

Internal and external comparisons

In order to determine the direction and degree of association between internal and external rater attributed scores for technical and non-technical performance, correlation analysis was completed using Spearman's rank correlation coefficient (r_s). In order to assess whether the internal raters were more/less/equivalently stringent to the external rater, the Wilcoxon signed rank test was used to determine the mean rank difference between external and internal attributed scores, again for technical and non-technical trainee performance.

Statistical analysis

All statistical analyses were completed using SPSS version 23.0 (IBM SPSS Statistics, IBM Corp., Armonk, NY, USA). All descriptive statistics are presented as median (IQR), unless otherwise specified.

5.4 Results

Participants

Fifteen cases were observed in the real operating room by an internal and external rater for both technical and non-technical performance. The internal rater was one of four General Surgeons (each observing a median of 3.5 procedures (range: 2-6) and the external rater was P.S. The internal raters consisted of three males and one female. The median clinical experience of the internal raters was 13 years (range: 1-33). Five residents (two PGY-2s, two PGY-4s and one PGY-5) were observed a median of 3 procedures (range: 1-5).

Correlation between internal and external attributed scores

For technical performance, a moderately positive association ($r_s = 0.618$, $p = 0.014$) was demonstrated between scores attributed by internal and external raters (Figure 7). For non-technical performance, internal and external raters demonstrated a strongly positive association between attributed scores ($r_s = 0.731$, $p = 0.002$) (Figure 8).

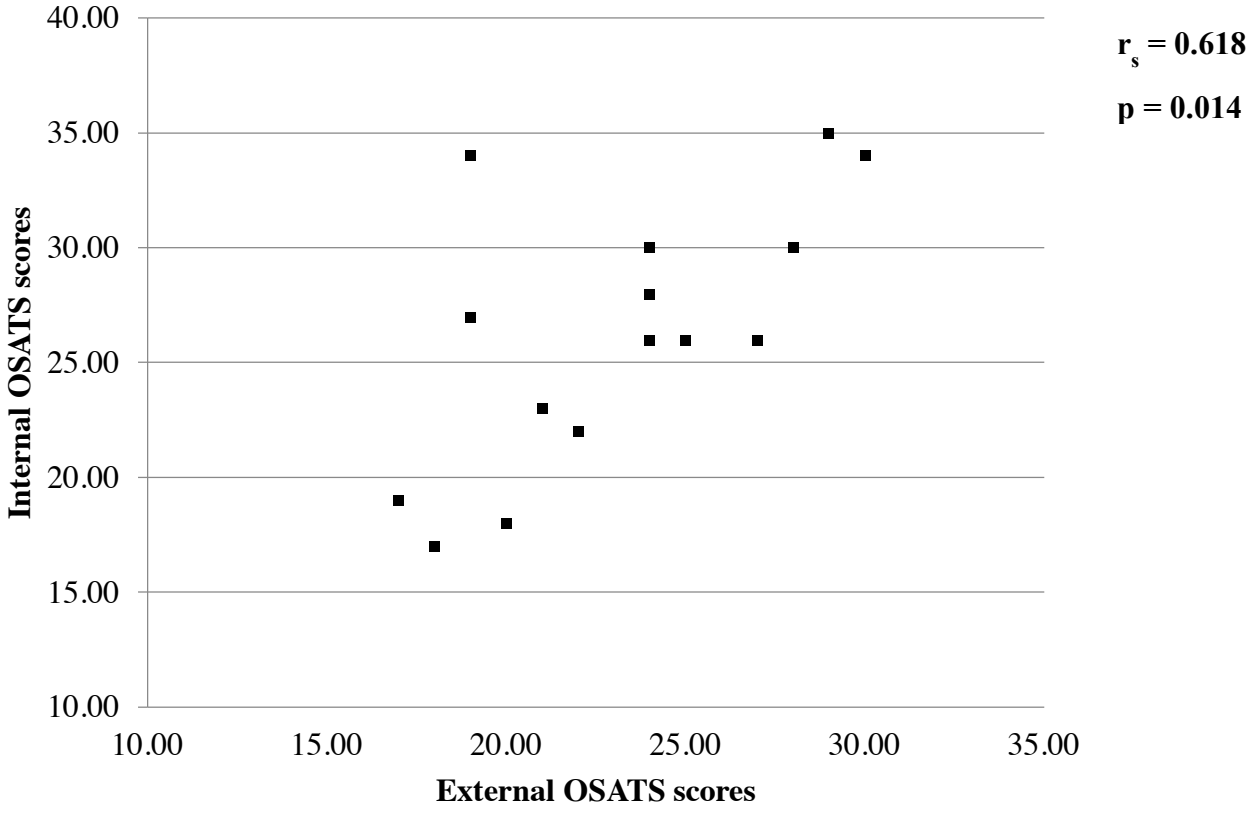

Figure 7. Correlation between external and internal rater attributed scores for technical performance.

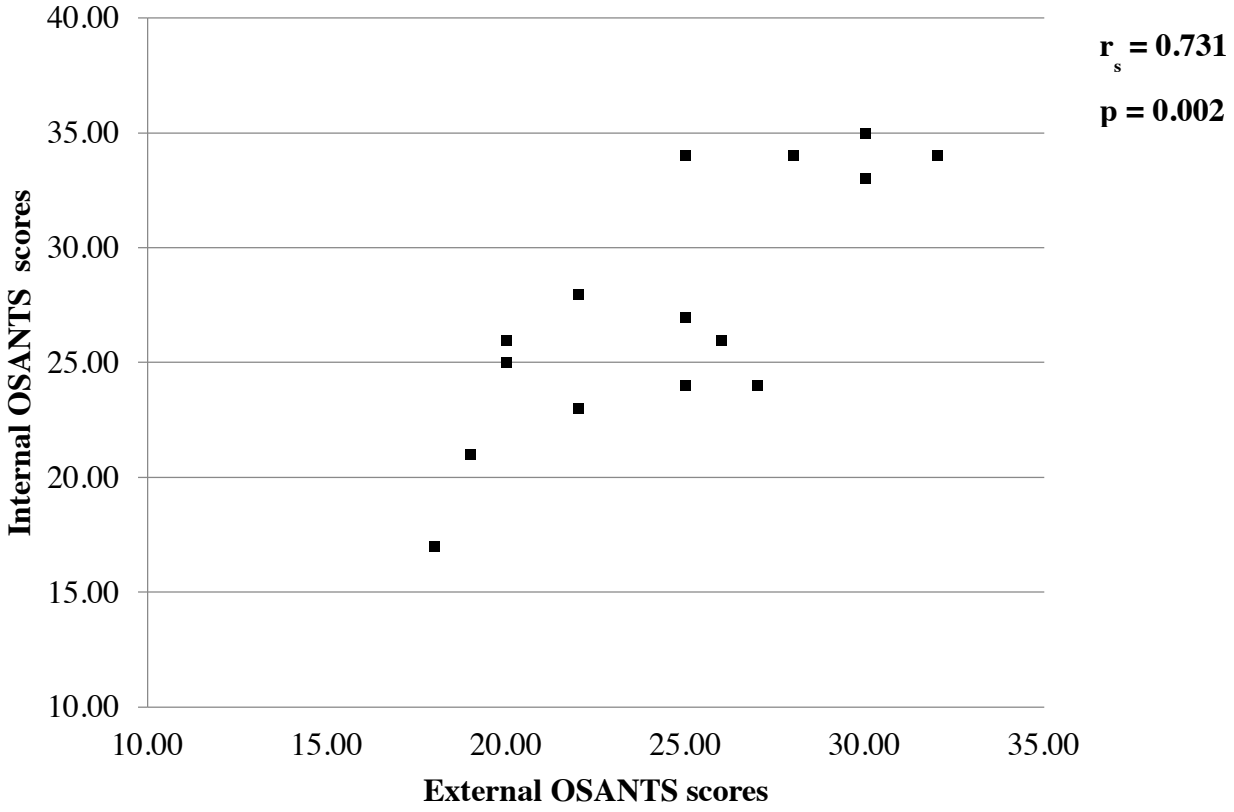

Figure 8. Correlation between external and internal rater attributed scores for non- technical performance.

Stringency of internal and external raters

For technical performance, the median external score was 24.0 (8.0), while the median internal score was 26.0 (8.0). In 11 cases, internal raters attributed a higher score per technical performance than the external rater. In three cases, the internal raters attributed a lower score per technical performance than the external rater. While in one case the internal and external scores were equal. These findings indicate that for technical performance, internal raters attributed higher scores to trainees than the external rater (mean rank 3.33 vs. 8.64) and this observed difference was found to be statistically significant (z-value = - 2.685, p = 0.007) (Table 9).

For non-technical performance, the median external score was 25.0 (8.0), while the median internal score was 26.0 (10.0). In 11 cases, internal raters attributed a higher score per non-technical performance than the external rater. In three cases, the internal raters attributed a lower score per non-technical performance than the external rater. While in one case the internal and external scores were equal. These findings also indicate that for non-technical performance, internal raters attributed higher scores to trainees than the external rater (mean rank 3.83 vs. 8.50) and this observed difference was found to be statistically significant (z-value = -2.583, p = 0.01) (Table 10).

Table 9. Difference in performance scores attributed per case and overall by internal and external raters for technical performance

Operative case	External score (OSATS)	Internal score (OSATS)	Difference in scores*
1	18	17	- 1
2	19	34	+ 15
3	19	27	+ 8
4	24	26	+ 2
5	20	18	- 2
6	22	22	0
7	27	26	- 1
8	24	28	+ 4
9	24	30	+ 6
10	21	23	+ 2
11	30	34	+ 4
12	29	35	+ 6
13	28	30	+ 2
14	17	19	+ 2
15	25	26	+ 1
Median (IQR)	**24.0 (8.0)**	**26.0 (8.0)**	**+ 2.00 *#**

*Difference in relation to external score (internal - external score)
Difference is significant (z-value = - 2.685, p = 0.007) (Wilcoxon signed rank test)

Table 10. Difference in performance scores attributed per case and overall by internal and external raters for non-technical performance

Operative case	**External score (OSANTS)**	**Internal score (OSANTS)**	**Difference in scores***
1	18	17	- 1
2	28	34	+ 6
3	20	26	+ 6
4	27	24	- 3
5	19	21	+ 2
6	22	23	+ 1
7	26	26	0
8	22	28	+ 6
9	25	34	+ 9
10	25	27	+ 2
11	32	34	+ 2
12	30	35	+ 5
13	30	33	+ 3
14	20	25	+ 5
15	25	24	- 1
Median (IQR)	**25.0 (8.0)**	**26.0 (10.0)**	**+ 1.0*[#]**

* Difference in relation to external score (internal - external score)
[#] Difference is significant (z-value = -2.583, p = 0.01) (Wilcoxon signed rank test)

5.5 Discussion

This study compared the ratings given by internal raters versus an external rater in terms of the technical and non-technical performance scores they attribute to trainees in the operating room during a laparoscopic cholecystectomy. The results indicate that although there is a moderate to strong positive correlation between internal and external rater attributed scores, the internal raters overall were less stringent than the external rater. This information is timely and useful to training programs and governing bodies looking to implement competency-based assessments, some of which will undoubtedly be summative in nature. Given the time, cost, resources and infrastructure required for competency-based education and assessment, the use of internal assessors could minimize and perhaps even circumvent some of these logistical issues (Del Bigio, 2007; Takahashi, 2011).

To the knowledge of the authors, there is a paucity of literature evaluating correlations between internal and external-type raters, both in the simulated setting and the operating room. The results presented here can, however, be compared to evaluations completed by different sets of raters – such as self-assessments compared to internal assessments and evaluations completed by staff and resident raters (both types of internal assessors) (Goldstein et al., 2014; Ward et al., 2003). They can also be compared to assessments completed by raters and compared to more formalized and standardized evaluations methods (L. S. Feldman et al., 2004; Reid et al., 2014). Comparing our results to these demonstrates similar findings, suggesting a moderately strong - to - strong positive correlation between the internal and external raters, in our case for both technical and non-technical performance assessments. The correlations in this study were similar or slightly stronger than evaluations completed by different sets of raters, as well as correlations previously completed comparing raters and more formalized evaluations (L. S. Feldman et al., 2004; Goldstein et al., 2014; Reid et al., 2014; Ward et al., 2003). This information could suggest that despite being intimately tied to the learning process, internal raters, once trained, can provide accurate assessments of technical and non-technical performance and this is likely suitable for formative-type assessments.

Despite these positive correlations for both types of performances, there was still a significant difference between the internal raters and external rater, suggesting that internal raters were less stringent than the external rater. The concept of leniency/stringency (a type of rater

error) in performance-based assessments is well known (Bartman, 2011; S. M. Downing, 2005; Myford & Wolfe, 2004). Previously documented reasons for rater leniency include 1) method of rater training, 2) the inexperience of the raters in number of cases observed – suggesting that an increase in cases observed leads to a more stringent rater, and 3) demographic characteristics of the raters – McManus *et al*. suggested that the ethnic background of raters in the United Kingdom resulted in a difference in performance ratings, while that study and others have not found such differences related to the gender or clinical experience of the raters (Bartman, 2011; Harasym, Woloschuk, & Cunning, 2008; McManus, Thompson, & Mollon, 2006). The implications of rater leniency, especially for those trainees that are around the pass/fail cut-point, where a small difference in attributed scores can lead to significant ramifications for both the trainee and training program are of utmost importance (Cizek & Bunch, 2007f; Harasym et al., 2008; Littlefield et al., 1991). This is particularly true for summative-type decisions (Boulet, McKinley, Whelan, & Hambleton, 2003). Therefore, attempting to align the internal raters with external raters relies directly on being able to address and modify the reasons that perpetuate rater leniency (Bartman, 2011; Harasym et al., 2008; McManus et al., 2006).

Possible methods to address the leniency of raters include: 1) rater training and re-assessment, 2) the use of rater pairs, and 3) statistical methods to mitigate rater stringency/leniency. There are multiple techniques used to train raters (M. Feldman et al., 2012). Despite some divergent studies, rater training has been shown to improve the accuracy and reliability of rater assessments (D. A. Cook, Dupras, Beckman, Thomas, & Pankratz, 2009; M. Feldman et al., 2012; Holmboe, Hawkins, & Huot, 2004; Woehr, 1994). With regards to rater training, what seems most important is that some type of training takes place, with the specified training regimen possibly reducing certain aspects of rater error (such as central tendency, leniency/stringency, halo effect etc.) (M. Feldman et al., 2012; Williams et al., 2003). For rater leniency in particular, RET and Frame of Reference (FOR) training have been previously shown to improve leniency, while PDT has been shown to worsen rater leniency (Williams et al., 2003; Woehr, 1994). The training regimen chosen to reduce rater leniency, however, also needs to be considered in the context of the other sources of rater error and how these various training paradigms work together to mitigate these sources (M. Feldman et al., 2012; Williams et al., 2003; Woehr, 1994). Rater training should also be an active process with ongoing re-assessment and re-calibration, to ensure that assessments continue to be reliable and accurate and to ensure

rater DRIFT (differential rater function over time) – a phenomenon whereby raters can become more or less lenient over time – does not occur (M. Feldman et al., 2012; McLaughlin, Ainslie, Coderre, Wright, & Violato, 2009; Williams et al., 2003). As reported by Bartman *et al*., another way to possibly lessen the effects of overly strict raters (in either direction) is to have rater pairs assess each performance rather than a single rater, acknowledging as they have, that this does carry with it many logistical issues (Bartman, 2011). Finally, the use of statistical methods to mitigate rater stringency/leniency may be most feasible. The premise relies on determining and possibly correcting for rater variance (error) within attributed performance scores (Harasym et al., 2008; C. Roberts, Rothnie, Zoanetti, & Crossley, 2010; C. Roberts, Zoanetti, & Rothnie, 2009). Many methods have been developed and they include item response theory (IRT) techniques such as Multi-Faceted Rasch Models (MFRM) (Harasym et al., 2008; C. Roberts et al., 2010; C. Roberts et al., 2009), the mean of the unsigned differences (M_uD) (Boulet et al., 2003) technique and the use of Generalizability (G) theory (Boulet et al., 2003; MacMillan, 2000; Myford & Wolfe, 2004).

In light of the findings in this study that internal raters are more lenient and previously published literature suggesting reasons for such leniency and possible techniques to address it, the question that remains is whether internal raters can be used in summative type assessments. Summative assessments are those from which decisions about trainees can be made and thus such assessments can have significant implications for all parties involved (Holmboe et al., 2010). Certain important features within summative type assessments include having a planned and appropriate structure to the development of the assessment, ensuring the results obtained are valid (and reliable) and, that the pass/fail criteria have been developed in a systematic manner (R.J. Mislevy, 2003; R.J. Mislevy, 2011; J. J. Norcini, 2003; Wass et al., 2001). Whether internal raters can be used within summative assessments falls under the category of obtaining valid (reliable) results. Although correlation coefficients do not exist to document what minimum threshold is required when comparing different sets of raters (i.e. internal to external), Downing has suggested that rater reliability (often measured using the intra class correlation coefficient [ICC]) of at least 0.80 for summative type assessments should be used (S. M. Downing, 2004). Although not documented directly for a Spearman's rank correlation coefficient, a Pearson product moment correlation coefficient, although close in value to an ICC, is likely an overly liberal reliability measure (Streiner, 2008). Furthermore, in order to ensure

valid (and reliable) scores have been attributed by the internal raters, evidence must be collected to support the interpretation of these results, in keeping with Messick's framework (S. M. Downing, 2003; Messick, 1995). In particular, the domains of response process, internal structure and consequences can be focused on and highlighted as has been done in this and future studies, in developing and collecting evidence for the use of summative type assessments where internal raters will be utilized (S. M. Downing, 2003; Messick, 1995). Finally, trainees should be observed by multiple raters over multiple occasions to ensure that the attributed scores are an accurate reflection of a trainee's 'true' performance (Crossley et al., 2007). Although, this study is the first step in the right direction, based on the results and the lack of other robust data to support or refute the use of internal raters who are intricately tied to resident training, the verdict on their use in summative assessments is still pending.

There are two potential limitations within this study. The first is whether the four internal raters used are representative of faculty raters across institutions with respect to the idea of rater leniency versus stringency presented in this study, also known as the dove/hawk effect, the latter term representing two extremes along a continuum (McManus et al., 2006). Given that the internal raters were seen as more lenient than the external rater, it is possible that the internal raters utilized in this study were inherently more lenient (doves) compared to internal raters in general (McManus et al., 2006). Although a possibility, this is unlikely, as research has previously documented that most raters are somewhere in the middle and that doves are rare and hawks even rarer (McManus et al., 2006). Additionally, a previous study has documented the median and range of staff surgeon ratings for technical performance and the results presented here fall within that range, lending credibility to the idea that the internal raters in this study are likely representative of internal raters in general (S. L. de Montbrun, Satterthwaite, L.M., Grantcharov, T.P., 2015). A second limitation in this study has to do with not being able to determine an inter-rater reliability between the internal raters. In this study, having multiple General Surgeons observe the same case was not feasible, as there can only be a single primary General Surgeon per case. If multiple internal raters could be utilized, an inter-rater reliability could be calculated and if acceptably high, would perhaps lend further support to the ability of internal raters as a group to provide reliable scores that may be used in summative assessments. Although not feasible within this study, the authors believe this was somewhat mitigated as multiple raters were used, each observing multiple cases composed of different residents, instead

of focusing on a single internal rater whose inherent bias may have influenced the overall results. The literature has also suggested that trained raters, regardless of their clinical standing (i.e. staff or residents), rate in a similar manner (Dedy, Szasz, et al., 2015).

Future work should focus on determining what correlation coefficients between different sets of raters are appropriate for use in summative-type assessments as well as setting performance standards for both technical and non-technical competence that delineate cut-off points – scores around which more care will need to be taken to ensure appropriate decisions regarding trainee competence are being made. Finally, more evidence needs to be collected in order to determine whether internal raters can be used for summative-type assessments.

5.6 Conclusion

This study compared the ratings given by internal raters versus an external rater in terms of the performance scores they attributed to trainees in the operating room during a laparoscopic cholecystectomy. Internal raters were found to moderately correlate to the external rater for technical performance and strongly correlate to the external rater for non-technical performance. Overall, the internal raters were less stringent than the external rater for both technical and non-technical performance. This information is important and timely as training programs and governing bodies look to implement competency-based assessments, which will bring with them many logistical issues, some of which may be minimized if internal raters are utilized (Del Bigio, 2007; Takahashi, 2011). Although promising, and likely appropriate for formative assessments, next steps include evaluating whether internal rater stringency can be aligned with that of the external raters, and subsequently whether internal raters can be utilized for summative type assessments.

CHAPTER 6

SETTING PERFORMANCE STANDARDS FOR TECHNICAL AND NON-TECHNICAL COMPETENCE IN GENERAL SURGERY

Chapter purpose

This study sets credible and reliable performance standards for technical and non-technical performance in the operating room. Trainees are also evaluated for their ability to meet both standards concurrently, and factors are examined that can predict at what point (years in training and case experience) trainees are likely to reach competence. This chapter addresses the limitations identified in section 1.9.8 of Chapter 1.

Chapter preface

The contents of this chapter have been accepted to the *Annals of Surgery* as: Setting Performance Standards for Technical and Non-Technical Competence in General Surgery. **P. Szasz**, M. Louridas, E.M. Bonrath, A.B. Fecso, B. Howe, A. Fehr, M.Ott, L.A. Mack, K.A. Harris, T.P. Grantcharov (2016)

6 Setting Performance Standards for Technical and Non-Technical Competence in General Surgery

6.1 Abstract

Background: Scores on performance assessments are difficult to interpret in the absence of established standards. The objectives of this study were to (1) create a technical and non-technical performance standard for the laparoscopic cholecystectomy, (2) assess the classification accuracy and (3) credibility of these standards, (4) determine a trainees' ability to meet both standards concurrently, and (5) delineate factors that predict standard acquisition.

Methods: Trained raters observed General Surgery residents performing laparoscopic cholecystectomies using the OSATS and the OSANTS instruments, while also providing a global competent/non-competent decision for each performance. The global decision was used to divide the trainees into two contrasting groups and the OSATS or OSANTS scores were graphed per group to determine the performance standard. Parametric statistics were used to determine classification accuracy and concurrent standard acquisition, while ROC curves were used to delineate predictive factors.

Results: Thirty-six trainees were observed 101 times. The technical standard was an OSATS of 21.04/35.00 and the non-technical standard an OSANTS of 22.49/35.00. Applying these standards, competent/non-competent trainees could be discriminated in 94% of technical and 95% of non-technical performances ($p<0.001$). A 21% discordance between technically and non-technically competent trainees was identified ($p<0.001$). ROC analysis demonstrated case experience and trainee level were both able to predict achieving the standards with an area under the curve (AUC) between 0.83 – 0.96 ($p<0.001$).

Conclusion: This study presents defensible standards for technical and non-technical performance. Such standards are imperative to implementing summative assessments into surgical training.

6.2 Introduction

There are multiple instruments available to assess trainee technical and non-technical performance in the operating room (Dedy, Szasz, et al., 2015; Mishra et al., 2009; Szasz et al., 2015; Yule, Flin, Paterson-Brown, Maran, et al., 2006). In the absence of performance standards however, the scores on these assessments stratify trainees in an arbitrary manner and provide little information regarding the attainment of an appropriate level of performance (Cizek & Bunch, 2007c; Koehler & Nicandri, 2013). This lack of established standards is especially pertinent as surgical education transitions towards a CBME training paradigm with the eventual implementation of milestones – in essence opportunities for summative type assessments, where evaluations are used to make decisions about a trainee's ability to progress (i.e. promotion, remediation) (Cizek & Bunch, 2007c; Cogbill, 2014; J. R. Frank, Snell, L.S., Sherbino, J, 2014; Holmboe et al., 2010; T. J. Wilkinson et al., 2001). In order to circumvent this shortcoming inherent to many of the assessments currently used in surgery, standard-setting methodologies from the medical education literature can be adopted and applied to set performance standards (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Nungester et al., 1991; Schindler et al., 2007). Within General Surgery specifically, these standards (whether related to technical or non-technical performance) can then provide defensible evidence to program directors, governing bodies and the trainees themselves whether an appropriate level of performance for a particular domain or procedure has been reached at such summative assessments (milestones) (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. R. Frank, Snell, L.S., Sherbino, J, 2014).

Standard-setting is a process whereby cut scores on examinations or assessments are created in a systematic and prescribed manner (Cizek, 1993; Cizek & Bunch, 2007f). These cut scores (standards) then divide the performance of the examinees into two categories: those that are competent and those that are non-competent for a particular domain or procedure (Cizek & Bunch, 2007f; J. J. Norcini, 2003). There are multiple standard-setting methodologies currently employed within medical education (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). While some focus on test content (test-centered) and are used for written examinations (i.e. the USMLE), others focus on the test takers (examinee-centered) and are used to assess actual performance (Nungester et al., 1991; T. J. Wilkinson et al., 2001). One examinee-centered approach shown to be valuable in assessing and stratifying medical trainees

during clinical encounters is the contrasting groups methodology (Jacobsen et al., 2015; Konge et al., 2012).

Thus, the purpose of this study was to develop credible and reliable performance standards in General Surgery using the contrasting groups methodology. Specifically, the objectives of this study were to (1) create a technical and non-technical performance standard for the laparoscopic cholecystectomy, (2) assess the classification accuracy (reliability) and (3) credibility (validity) of these newly created performance standards, (4) determine a trainees' ability to meet the two performance standards concurrently and (5) determine which trainee factors predict standard acquisition.

6.3 Methods

Ethics

The research ethics boards at all participating institutions approved this study.

Program and resident participation

This was a multi-institution study involving three Canadian General Surgery programs: Western University, the University of Calgary and the University of Toronto. Within the three participating programs, all levels of residents, PGY 1-5, were eligible to participate.

Operative procedure

Based on previous work identifying procedures for milestone assessments, the laparoscopic cholecystectomy was chosen for this study (Szasz, Louridas, de Montbrun, Harris, & Grantcharov, 2016).

Data collection and assessment instruments

All data collection occurred in the operating room. The trainees' technical performance of a laparoscopic cholecystectomy was recorded using the intra-abdominal camera feed off the operating room computer and evaluated objectively by blinded expert raters at a later date via the OSATS (J. A. Martin et al., 1997) rating instrument (Appendix 3). The trainees' non-technical performance during the laparoscopic cholecystectomy was evaluated by expert raters in real-time

via the OSANTS (Dedy, Szasz, et al., 2015) rating instrument (Appendix 4). Each rating instrument is composed of seven categories, ranked on a five-point Likert scale with a maximum score of 35 (Dedy, Szasz, et al., 2015; J. A. Martin et al., 1997). In the instance that case takeover by the staff surgeon occurred during an observed procedure, that portion was not assessed by the expert raters.

The OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) instruments meet many of the criteria within Messick's conceptual framework of validity – providing support to use these assessment results (Dedy, Szasz, et al., 2015; S. M. Downing, 2003; Ghaderi et al., 2015; J. A. Martin et al., 1997; Messick, 1995; Reznick et al., 1997). While in a more traditional sense, the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) instruments have previously demonstrated construct validity, internal consistency and inter-rater reliability (Dedy, Szasz, et al., 2015; J. A. Martin et al., 1997; Reznick et al., 1997).

Raters and rater training

The raters in this study all had experience in surgical education and assessment. Prior to undertaking any ratings, all raters were formally trained in the application of the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) instruments using FOR and PDT approaches (M. Feldman et al., 2012). Initially, the rating instrument (either the OSATS (J. A. Martin et al., 1997) or OSANTS (Dedy, Szasz, et al., 2015)) was introduced to the raters and each scale item was fully explained. Subsequently, the raters observed independently a set of five real laparoscopic cholecystectomy videos for technical performance training (or five simulated scenario videos for non-technical training). These cases were then viewed and discussed in a group setting led by previously trained experts, focusing on exemplars of good and bad performance markers, as well as demonstrable examples for each scale item. Consequently, a second set of three real laparoscopic cholecystectomy videos for technical performance training (or three real-time observations in the operating room for non-technical training) were assessed and reviewed. The inter-rater reliability was found to be excellent with an ICC of 0.88 for OSATS (J. A. Martin et al., 1997) and between 0.83 to 0.96 for OSANTS (Dedy, Szasz, et al., 2015) (S. M. Downing, 2004).

Creation of the technical and non-technical performance standard

To create performance standards using the contrasting groups methodology, the trainees' performance was first scored by the raters using either the OSATS (J. A. Martin et al., 1997) (for technical performance) or OSANTS (Dedy, Szasz, et al., 2015) (for non-technical performance) rating instruments (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Livingtson, 1982). Next, the trainees were divided into competent and non-competent groups for technical and non-technical performance separately (each type of performance giving rise to two contrasting groups) based on the same raters' assessment of the trainees' overall global performance (providing an overall competent/non-competent classification) (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Livingtson, 1982). In the context of this study, competence was defined as the ability to safely and effectively perform a laparoscopic cholecystectomy (Szasz et al., 2015). When multiple raters were utilized per assessment, majority consensus was used to arrive at the competent/non-competent decision and the average of the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) scores were utilized (J. J. Norcini, Shea, JA, 1997). For the competent/non-competent classification, three outcomes were possible: 1) complete agreement amongst the judges, 2) majority agreement amongst the judges, and 3) a split classification amongst the judges. In instances 1) and 2) no further discussions were required. In instance 3) a discussion occurred between judges (until a majority was reached) where the reasoning for assigning the trainee based on their performance to either the competent or non-competent group was provided with specific examples. This discussion was based on the way the raters were trained using the Frame of Reference (FOR) and Performance Dimension Training (PDT) approaches (where good and bad performance markers were discussed) (M. Feldman et al., 2012). The overarching question that was returned to during the discussion was whether the trainee in question could safely and effectively complete a laparoscopic cholecystectomy (Szasz et al., 2015).The score distributions from the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) rating scale for each contrasting group were graphed and the means and standard deviations calculated. The cut score (performance standard) was then set where the distributions intersect (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Livingtson, 1982).

Classification accuracy of the performance standards

The classification accuracy (reliability) of the performance standards was determined by comparing each technical and non-technical, competent/non-competent classification based on the cut score (performance standard) to the initial competent/non-competent classification made by the expert raters, using the chi-square square test for independence (χ^2) and percent agreement.

Credibility of the performance standards

The validity (credibility) in the context of performance standards depends on collecting evidence to support their use in a particular setting (J. J. Norcini, Shea, JA, 1997). This is accomplished by adhering to three categories within Norcini's framework, including (1) appropriately selecting the standard setters, (2) utilizing sound methodologies, and (3) creating standards that are realistic (J. J. Norcini, Shea, JA, 1997).

Concurrent acquisition of the performance standards

The ability of trainees to achieve both the technical and non-technical performance standard during the same laparoscopic cholecystectomy was determined using the χ^2 test and percent agreement.

Factors that predict achieving the performance standards

ROC curves were used to identify trainee factors that predict standard acquisition. The sensitivity (competent trainees who met the standard) or true positive rate (TPR) and 1-specificity (non-competent trainees who met the standard) or false positive rate (FPR) for various PGY levels and case experiences was calculated determining the point (PGY level and case number) where both sensitivity and specificity were maximized, suggesting an appropriate tradeoff between the TPR and true negative rate (TNR) (Fraser et al., 2003). The area under the curve (AUC) represents the probability of accurately discriminating (based on knowing PGY level or case number) whether a General Surgery trainee would be able to meet the performance standards.

Statistical analysis

Statistical analyses were calculated using SPSS version 22.0 (IBM SPSS Statistics, IBM Corp., Armonk, NY, USA).

6.4 Results

Participating residents and expert raters

Thirty-six trainees were observed in the real operating room during a laparoscopic cholecystectomy, providing 101 technical and 100 non-technical performance assessments (on one occasion non-technical performance was not recorded). There were 23 males, 13 females and the median age was 29.50 (range: 25-42). The median number of performance assessments per trainee was 3 (range: 1-9). All PGY levels were represented (six PGY 1s, four PGY 2s, seven PGY 3s, nine PGY 4s, and ten PGY 5s). Trainees' previous experience varied from less than 10 to more than 100 laparoscopic cholecystectomies (eight trainees < 10, five: 11-25, seven: 26-50, three: 51-75, one: 76-99 and, eleven ≥ 100, data was unavailable for one trainee).

For technical performance, six trained raters were used in total with a median observation of 29.50 cases (range: 24-101) providing 242 performance assessments. For technical performance, the competent/non-competent classification occurred 101 times. In 68 of these instances two or more raters were involved. Of the 68 classifications with two or more raters (34 cases were assessed by two raters, 7 cases by three raters, 15 cases by four raters, and 12 cases by five raters) complete classification agreement occurred in 42 instances, majority consensus was met in 12 cases and there was split agreement in 14 cases, which underwent discussion until a majority was achieved. Overall, percent agreement for all technical classifications was 84.2% and the inter-rater reliability of the OSATS (J. A. Martin et al., 1997) scores was found to be excellent with an intra-class correlation coefficient (ICC) of 0.88 (S. M. Downing, 2004; McHugh, 2012).

For non-technical performance, four trained raters were used in total with a median observation of 19.50 cases (range: 4-100) providing 143 performance assessments. For non-technical performance, the competent/non-competent classification occurred 100 times. In 37 of these instances two or more raters were involved (31 cases were assessed by two raters and, 6

were assessed by three raters). This was a smaller number than for technical performance, given the feasibility and acceptability of having multiple raters in the real OR at one time. Complete classification agreement occurred in 33 instances and there was split agreement in 4 cases, which underwent discussion until a majority was achieved. Overall, percent agreement for all non-technical classifications was 95% and the inter-rater reliability of the OSANTS (Dedy, Szasz, et al., 2015) scores was found to be excellent with an ICC between 0.83 to 0.96 (S. M. Downing, 2004; McHugh, 2012).

Performance standards

The technical performance mean score of the trainees who comprised the non-competent group was 17.60 (SD 2.30), while the mean score for the trainees who comprised the competent group was 24.78 (SD 2.71) ($p<0.001$). The technical performance standard was found to be an OSATS(J. A. Martin et al., 1997) score of 21.04/35.00 (Figure 9A).

The non-technical performance mean score of the trainees who comprised the non-competent group was 19.27 (SD 1.63), while the mean score for the trainees who comprised the competent group was 27.33 (SD 2.91) ($p<0.001$). The non-technical performance standard was found to be an OSANTS(Dedy, Szasz, et al., 2015) score of 22.49/35.00 (Figure 9B).

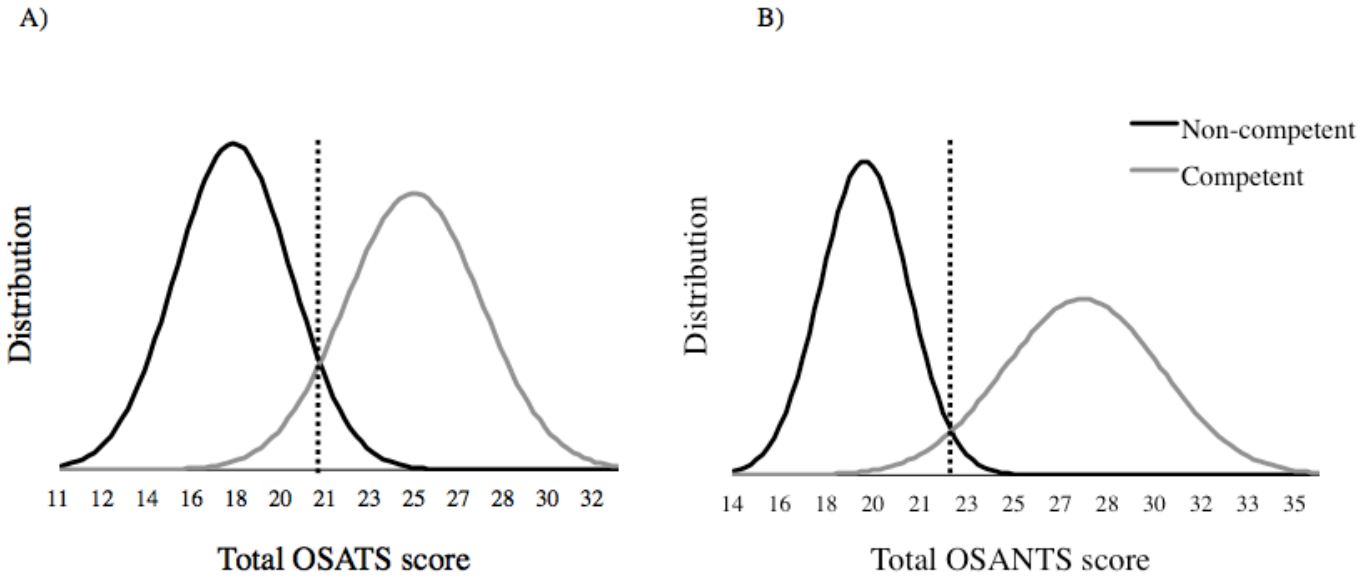

Figure 9. Contrasting group curves for technical and non-technical performance. (A) Technical performance cut score was found to be an OSATS of 21.04/35 ($p<0.001$). (B) Non-technical performance cut score was found to be an OSANTS of 22.49/35 ($p<0.001$).

Classification accuracy of the standards

The classification accuracy of the technical performance standard was found to be 94% (χ^2=78.46, p<0.001) (Table 11). Two trainees were found to be competent based on the cut-score, but not the global decision (false positives). While four trainees were found to be non-competent based on the cut-score, but not the global decision (false negatives).

The classification accuracy of the non-technical performance standard was found to be 95% (χ^2=80.13, p<0.001) (Table 12). Five trainees were found to be non-competent based on the cut-score, but not the global decision (false negatives). There were no false positives for non-technical performance.

Table 11. Classification accuracy (reliability) of the technical performance standard

		Classification based on cut score (performance standard)	
	Competent	**Non-competent**	**Total**
Classification based on global trainee performance			
Competent	51 (50.5%)	4 (4.0%)	55 (54.5%)
Non-competent	2 (2.0%)	44 (43.6%)	46 (45.5%)
Total	53 (52.5%)	48 (47.5%)	101 (100%)

Chi Square χ^2 = 78.46, p < 0.001

Table 12. Classification accuracy (reliability) of the non-technical performance standard

		Classification based on cut score (performance standard)		
		Competent	**Non-competent**	**Total**
Classification based on global trainee performance	**Competent**	63 (63.0%)	5 (5.0%)	68 (68.0%)
	Non-competent	0 (0%)	32 (32.0%)	32 (32.0%)
	Total	63 (63.0%)	37 (37.0%)	100 (100%)

Chi Square $\chi^2 = 80.13$, $p < 0.001$

Credibility of the standards

This section demonstrates with examples how each category within Norcini's framework was adhered to (J. J. Norcini, Shea, JA, 1997). (1) Appropriately selecting the standard setters – several trained judges were utilized for the technical and non-technical performance assessments, they varied in their qualifications (some were staff surgeons while others residents), each engaged in the profession under assessment and all were knowledgeable with regards to the content being assessed and the consequences of their decisions (R. L. Brennan, Lockwood, R.E, 1980; J. J. Norcini, Shea, JA, 1997). (2) Utilizing sound methodologies – the use of an absolute standard (the contrasting groups methodology) (J. J. Norcini, Shea, JA, 1997). Due diligence in the process of setting standards was demonstrated by the extent of rater training, the number of assessments completed by each rater, and the use of standards which are supported by research in similar fields (Jelovsek et al., 2010; J. J. Norcini, Shea, JA, 1997). (3) Creating standards that are realistic – such as the demonstration (in the sections to follow) that a relationship exists between meeting the performance standard and external markers of competence as evidenced by PGY level and previous case experience (J. J. Norcini, Shea, JA, 1997).

Concurrent acquisition of the standards

There was a 21% discordance (χ^2=34.86, p<0.001) when trainee technical and non-technical performances were assessed for the same laparoscopic cholecystectomy (Table 13). Sixteen trainees were able to achieve the non-technical standard, but not the technical standard, while five trainees reached the technical standard but not the non-technical standard.

Table 13. Concurrent achievement of the technical and non-technical performance standard by trainees during the same performance assessment

		Achievement of technical performance standard	
	Competent	**Non-competent**	**Total**
Achievement of non-technical performance standard			
Competent	47 (47.0%)	16 (16.0%)	63 (63.0%)
Non-competent	5 (5.0%)	32 (32.0%)	37 (37.0%)
Total	52 (52.0%)	48 (48.0%)	100 (100%)

Chi Square $\chi^2 = 34.86$, $p < 0.001$

Predictive factors

For technical performance, ROC analyses demonstrated that case experience (AUC: 0.83 [95% CI, 0.68 - 0.97]; p<0.001) and PGY level (AUC: 0.85 [95% CI, 0.72 – 0.98]; p<0.001) were equally able to predict standard acquisition (Figure 10) and that one was not statistically better than the other in doing so (p=0.86). The case experience where sensitivity (77%) and specificity (72%) were maximized was found to be 46.50 laparoscopic cholecystectomies. Conversely, the PGY level where the sensitivity (82%) and specificity (74%) were maximized was 3.5 years of training.

For non-technical performance, case experience (AUC: 0.93 [95% CI, 0.85 – 1.00]; p<0.001) and PGY level (AUC: 0.96 [95% CI, 0.91- 1.00]; p<0.001) were again equally able to predict standard acquisition (Figure 10) and there was no statistically significant difference between the two (p=0.57). The case experience where sensitivity (91%) and specificity (92%) were maximized was found to be 35 laparoscopic cholecystectomies, while the PGY level where the sensitivity (86%) and specificity (100%) were maximized was also 3.5 years.

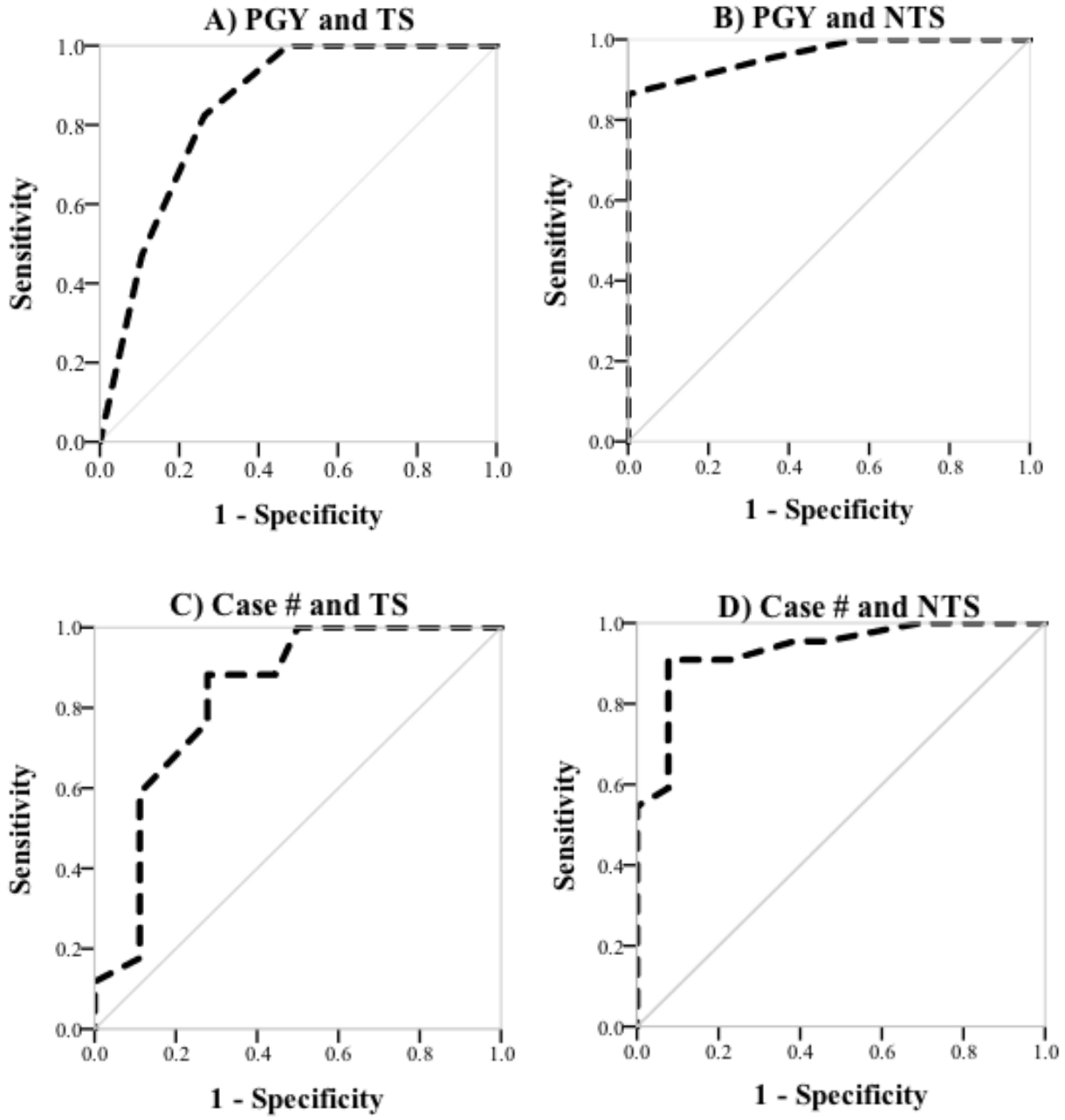

Figure 10. ROC curve analysis for the ability of trainee PGY level and case experience to predict attaining the technical (A and C) and non-technical (B and D) performance standards. For technical performance both A) PGY level (AUC: 0.85 [95% CI, 0.72 – 0.98]; p<0.001) and C) case expereince (AUC: 0.83 [95% CI, 0.68 - 0.97]; p<0.001) were able to predict acquisition of the standard and the difference between them was not statistically significant (p=0.86). For non-technical performance both B) PGY level (AUC: 0.96 [95% CI, 0.91- 1.00]; p<0.001)) and D) case experience (AUC: 0.93 [95% CI, 0.85 – 1.00]; p<0.001) were able to predict acquisition of the standard and this difference again was not statistically significant (p=0.57). ROC – receiver operator characteristic, PGY – postgraduate year, TS – technical standard, NTS – non-technical standard, CI – confidence interval.

6.5 Discussion

The recent focus and shift towards a CBME paradigm within surgical education requires a parallel shift in assessment practices (Holmboe et al., 2010). Although formative assessments will maintain their place as the cornerstones within this CBME framework, summative assessments are required to determine whether an appropriate level of performance on a particular domain or task has been reached (C. Carraccio et al., 2002; Govaerts et al., 2007; Holmboe et al., 2010). This is particularly true, as CBME will focus more on the demonstration of competence and less on the duration of training, lending little support for time as a surrogate marker of competence, as it once was (J. R. Frank, Snell, L.S., Sherbino, J, 2014). Within this CBME paradigm, the various milestones, the specifics of which are currently in development, can serve as a blueprint for such summative assessments (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Green et al., 2009; Holmboe et al., 2010).

Once summative assessments have been implemented, the next important feature is to ensure their results are based in systematically set performance standards rather than arbitrarily and that the standards display classification accuracy (Cizek & Bunch, 2007c; Koehler & Nicandri, 2013). There is a long history of using such standard-setting methodologies in other areas of education and medicine (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; McClarty, 2013; J. J. Norcini, 2003). Standards using the contrasting groups methodology have been set for procedures such as the bronchoscopy, phacoemulsification, colonoscopy and vaginal hysterectomy, demonstrating very reasonable classification accuracy between performers of various levels, with the false positive and false negative rates ranging between 0 to 19% (Jelovsek et al., 2010; Konge et al., 2012; Preisler et al., 2015; Thomsen, Kiilgaard, Kjaerbo, la Cour, & Konge, 2015). The ability of these standards to appropriately discriminate trainees into those that are competent and those that are not are comparable to the results presented in our study. All of these studies, however, either assessed trainees in a simulated setting, or used external markers (i.e. novice versus expert, total number of procedures) to initially divide participants into the two contrasting groups (Jelovsek et al., 2010; Konge et al., 2012; Preisler et al., 2015; Thomsen et al., 2015). To the knowledge of the authors, only one study has assessed trainees in a non-simulated setting while also using judgment of actual trainee performance to divide participants into each of the contrasting groups (Livingtson, 1982; Sedlack et al., 2016). That particular study set standards for diagnostic colonoscopies evaluating the competence of

gastroenterology fellows (Sedlack et al., 2016). Within CBME, summative assessments will likely occur in a work place setting and use judgment of actual performance to initially divide trainees instead of external factors (C. Carraccio et al., 2002; Holmboe et al., 2010). As a result, studies like the one presented in this paper are essential, especially for surgical disciplines where such evidence is currently lacking.

In addition to setting standards and ensuring they possess classification accuracy during summative assessments, the next component is to demonstrate the standards are also credible. Credibility in the context of performance standards is based on work proposed by Norcini, who suggested that rather than validating a standard (establishing its correctness), it is more fitting to gather evidence that supports the use of that standard for a specific purpose (J. J. Norcini, Shea, JA, 1997). Norcini suggests that care should be taken to utilize a varied group of expert judges, create absolute standards and demonstrate that a relationship exists between the trainees who meet the standard and other markers that evaluate the same construct under study (J. J. Norcini, Shea, JA, 1997). Although there is no minimum amount of evidence that must be collected to achieve credibility, each category presents issues for consideration when developing and subsequently utilizing performance standards (J. J. Norcini, Shea, JA, 1997). Within this study, the development of both the technical and non-technical standards adhered to various levels of evidence within each category, lending support for use of these standards in differentiating technically or non-technically competent trainees for the laparoscopic cholecystectomy. Compared to other studies utilizing performance standards as described above, Jelovsek *et al*. also utilized Norcini's framework to collect evidence that supported the use of their standards, while Thomsen *et al*. and Sedlack *et al*. collected evidence regarding the validity of their scores, which were then used to create their performance standards (S. M. Downing, 2003; Ghaderi et al., 2015; Jelovsek et al., 2010; Messick, 1995; J. J. Norcini, Shea, JA, 1997; Sedlack et al., 2016; Thomsen et al., 2015).

Once credible and reliable standards have been created for use during summative assessments, the purpose of the evaluation must also be considered. In this study, false-negative decisions (trainees who are truly competent, but did not meet the performance standard) and false-positive decisions (trainees who are truly non-competent, but did meet the performance standard) were deemed equivalent. This was because the purpose of this particular assessment is yet to be defined. The decisions and subsequent repercussions from all summative assessments,

however, are not always equal (Hambleton, 1978; Livingtson, 1982). Take for example, summative evaluations completed during the junior years of training. Although high-stakes in nature, as decisions regarding trainee promotion may result, false negative misclassifications of trainees at this time are likely more detrimental to the trainee and training program than false positive misclassifications. This is especially true if a lengthy and expensive remediation program is required and there will be subsequent summative assessment opportunities for the false positive trainee (to further delineate their true competence). In contrast, summative evaluations completed at the certification time-point, where false positive misclassifications of trainees are likely more detrimental to the trainee and the public as the trainees become certified to undertake independent practice, as opposed to false negative decisions, which will require further trainee re-assessment. As a result, using the contrasting groups methodology, techniques based in statistical decision theory have been developed to manipulate the cut score in order to minimize either the false negative or false positive decision, depending on the purpose of the test (Hambleton, 1978; Livingtson, 1982). Using multiple assessors to evaluate trainees on multiple occasions is also important to ensure that observed performances closely reflect trainee 'true' performance, at the time of such summative assessments (Crossley et al., 2011; Crossley et al., 2007; Marriott et al., 2011). Although not all trainees in this study were watched multiple times, the purpose of this work was to create performance standards utilizing multiple trainees, raters and number of cases (as many as were feasible) – that in the future may be used for summative purposes. Moving forward if a training program or governing body were to implement these standards into training, it is at this time that multiple observations of a trainee are necessary in order to ensure trainee 'true' performance has been captured in order to avoid misclassifications at such assessments (Crossley et al., 2011; Crossley et al., 2007; Marriott et al., 2011).

Based on the performance standards presented here, a trainee would need to complete 46.50 laparoscopic cholecystectomies or 3.5 years of training in order to be deemed technically competent, and 35 laparoscopic cholecystectomies or 3.5 years of training in order to be deemed non-technically competent. These case numbers are realistic and fall below those reported by trainees in the U.S. and Europe for the laparoscopic cholecystectomy during a five year training period (R. H. Bell, Jr. et al., 2009; Carlsen et al., 2014; R. S. Chung, 2005). The intriguing point about these case numbers is that as CBME moves away from time-based training, they can represent the minimum range of procedures expected for the laparoscopic cholecystectomy to

ensure trainee competence (R. H. Bell, Jr. et al., 2009; Carlsen et al., 2014; R. S. Chung, 2005). This information can then go on to inform change within General Surgery training based on evidence, and bridge the knowledge gap that currently exists between the time-focused and the competence-focused training frameworks.

Finally, the findings in this study – that being technically competent did not imply non-technical competence or vice versa – are also key. Although technical and non-technical performance are different constructs, it is often anticipated that as trainees progress through training focusing on technical performance, they will acquire non-technical skills as well ("Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0," 2012). Adding the findings from this study to literature demonstrating that non-technical performance impacts patient safety provides further evidence for non-technical training during residency and not solely relying on trainees to independently acquire these skills ("Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0," 2012; Greenberg et al., 2007; Mishra et al., 2008). The information presented here may also suggest that non-technical skills are likely compensatory between different procedures (i.e. situational awareness learned during a laparoscopic cholecystectomy can be applied to increasing one's situation awareness during other procedures), as case numbers were lower for reaching the non-technical standard and there were fewer trainees failing to do so.

There are a few limitations that must be acknowledged within this study. The first relates to the inability to perform blinded non-technical performance assessments, given that the raters had to be present in real-time in the operating room. This was mitigated as much as possible by having multiple raters present in the operating room at the University of Toronto, where there was a potential (although unlikely) for clinical familiarity between the raters and trainees, and this demonstrated excellent inter-rater reliability. While at Western University and the University of Calgary this was mitigated, as the rater had no clinical familiarity with the trainees being assessed. The second limitation relates to the self-reported nature of the demographic data that was collected. Although care was taken to ensure the data acquired from trainees was as accurate as possible, trainees may have over or underinflated their case experience. Nonetheless, as demographic data in education studies is almost exclusively self-reported, this is an inherent shortcoming of most education research, and is not unique to this study.

6.6 Conclusion

This study establishes that credible and reliable technical and non-technical performance standards can be set in General Surgery for the laparoscopic cholecystectomy. It was further recognized that trainee characteristics including PGY level and previous case experience were equally capable as predictive factors, detailing at which level and case number trainees would likely acquire either performance standard. Lastly, during an operative case, trainees could be seen as competent in one performance domain, but not another.

These findings are timely as surgical education finds itself in a transitional state, with a real need to implement summative assessments that will determine whether appropriate levels of performance have been reached. The standards identified here provide defensible judgments during such summative assessments for the laparoscopic cholecystectomy and provide a starting point for the creation of additional assessments. While the information provided by the predictive factors and the inability to achieve concurrent competence by some trainees will help inform change within General Surgery training as the transition towards CBME occurs.

CHAPTER 7

GENERAL DISCUSSION

Chapter purpose

This chapter ties together the knowledge gaps currently present in surgical education identified in Chapter 1 and the findings in this thesis that aim to answer or provide evidence to fill these gaps. This general discussion is in addition to the study specific discussions contained in each chapter.

7 General Discussion

7.1 Summary of studies

This thesis and the studies within it, aimed to collect evidence that will help bridge many of the knowledge gaps currently present in surgical education, in order to support the transition towards CBME in General Surgery. Each study is briefly summarized below. This is followed by specific discussions focusing on the main topic areas covered within the thesis, the novelty of the study findings and their relationship to the currently available literature.

The first study (Chapter 1) focused on Aim #1 and described a systematic review of the literature evaluating the methods by which technical competence is assessed in surgical trainees, the psychometric properties of these methods and whether standard setting methodologies have been utilized in the included studies (Szasz et al., 2015). A unified definition of technical competence was also evaluated (Szasz et al., 2015). This study demonstrated that there are five methods by which technical competence is currently assessed in surgical trainees, with the majority of these having previously demonstrated validity and reliability, or in a more modern sense having met criteria within Messick's conceptual framework of validity (Messick, 1995). There seems to be a void in the use of standard setting methodologies in the included studies, with only four having some aspect of standard setting incorporated into them in the specialties of Vascular Surgery, Obstetrics & Gynecology and General Surgery although the ones in General Surgery occurred in the simulated setting (J. D. Beard et al., 2005; Jelovsek et al., 2010; Sroka et al., 2010; Swanstrom et al., 2006). This study also provided a unified definition of technical competence – the ability to safely and effectively complete an operation or procedure independently (Szasz et al., 2015).

The second study (Chapter 3) focused on Aim #2 and delineated an international perspective regarding technical competence assessments during all stages of surgical training, namely: selection into training, in-training progression and certification (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). The current and ideal (including evaluation instruments and locations of assessments) technical performance assessment practices were also sought as were real and/or perceived barriers to their adoption. Accruing the views of eight international education directorates (or their equivalents) this study demonstrated that for

the most part, technical performance is assessed at the in-training time point in either the simulated or workplace setting utilizing assessment instruments (global rating and task specific scales), with limited technical performance assessments occurring at the selection and certification time points. Ideal in-training practices include the assessments instruments these jurisdictions are already using and ITERs, with the workplace setting being favored over the simulated setting. At the selection time point ideal practices would incorporate basic or intermediate tasks in the simulated setting, while for certification ideal practices would include using again either the assessment instruments these jurisdictions are currently using or ITERs, with the workplace favored over the simulated setting. Finally, barriers to the adoption of such assessment practices include financial limitations for the in-training time point, a lack of evidence for their use during selection and a combination of these plus lack of faculty expertise for the certification time point.

The third study (Chapter 4) focused on Aim #3 and depicted a contemporary training and assessment model that takes recent changes to surgical training into consideration and can help form the foundation of CBME moving forward in General Surgery (Szasz et al., 2016). Specifically, a training model for operative procedures and tasks for junior and senior level trainees was outlined, as was an assessment model – delineating the procedures that can be used for technical milestone assessments, to determine whether junior level trainees can progress to senior level trainees (Szasz et al., 2016). This study demonstrated that there are 101 procedures and tasks that Canadian General Surgery program directors believe trainees should be competent to perform (Szasz et al., 2016). These include 24 specific to junior level trainees, 68 specific to senior level trainees and nine common to all trainees (Szasz et al., 2016). Four procedures were designated for use as milestone assessments (Szasz et al., 2016).

The fourth study (Chapter 5) focused on Aim #4 and determined whether staff surgeons who are intimately tied to training residents, can provide accurate performance assessments. This was accomplished by comparing the ratings provided by internal raters (staff surgeons) and an expert external rater (no relationship to the residents), in terms of the technical and non-technical performance scores they attribute to trainees in the operating room during a laparoscopic cholecystectomy. This study demonstrated a moderate association between internal and external raters for technical performance and strong association for non-technical performance. Furthermore, it was demonstrated that overall, internal raters were less stringent for both

technical and non-technical performance assessments compared to their external rater counterpart.

The fifth and final study (Chapter 6) focused on Aim #5 to construct performance standards that can delineate competent trainees from non-competent trainees. Specifically, credible and reliable technical and non-technical performance standards for the laparoscopic cholecystectomy were sought. As was information regarding whether trainees were able to meet both standards concurrently, and which trainee factors could predict standard acquisition. This study demonstrated that technical and non-technical performance standards could be created using the contrasting groups methodology and that these standards were credible and reliable. It was also demonstrated that there was a substantial (21%) discordance between trainees who were competent technically and non-technically or vice versa. Finally, years of training and case experience were equally able to predict attaining the performance standards.

7.2 Assessment of technical performance

7.2.1 Current methods employed

The traditional training and assessment frameworks used to educate surgical residents, focused almost exclusively on knowledge and judgment ("Curriculum outline for General Surgery Residency," 2015-2016; "Objectives of Training in the Specialty of General Surgery," 2010). The focus on these two domains is also reflected in the RCPSC certification examination completed at the end of training and the ABS licensing examination ("ABS Booklet of Information Surgery," 2015; "Format of the Comprehensive Objective Examination in General Surgery," 2014). With the implementation of CBME, this focus on knowledge and judgment almost exclusively is slowly changing both in regards to training and assessment (from both a formative and summative perspective). Now a variety of competencies/domains, all of which together contribute to the creation of a fit independent practitioner are seen as important ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014; J. R. Frank, Snell. L., Sherbino, J, 2015; Holmboe et al., 2015). A key concept within CBME centers on the need for assessments to occur in the workplace, rather than the simulation laboratory using appropriate methods (Hawkins et al., 2015; Holmboe et al., 2010). As a result, although the methods, by which non-technical performance is assessed, have been previously described, no such information has been collected

for technical performance, with a real need to delineate the methods by which technical competence (irrespective of whether it is done for formative or summative reasons) is currently assessed in surgical trainees (Dedy, Bonrath, Zevin, & Grantcharov, 2013; J. Scott et al., 2016; Yule, Flin, Paterson-Brown, & Maran, 2006). Thus, this lack of information relating to technical performance assessments formed the basis for the systematic review summarized above.

The implications of this study's findings for CBME are numerous.

Firstly, we provide a unified definition for medical/surgical educators as to the definition of technical competence and how this differs from technical skill, operative competence, surgical competence and proficiency (Szasz et al., 2015). This lack of a proper definition and/or inappropriate use has been documented in the literature to be a major barrier to the incorporation of performance assessments (Aggarwal, 2010; C. L. Carraccio & Englander, 2013). Although some work has previously been completed evaluating specific performance definitions, this work has focused mainly on dictionary designations, it stopped short of discriminating between the various definitions, and it is outdated having been accrued prior to the real transition towards CBME (Satava, Cuschieri, Hamdorf, & Metrics for Objective Assessment of Surgical Skills, 2003). Conversely, other definitions have been overly generic and intangible (J. R. Frank, Snell, et al., 2010). While still others have focused on the five stages of skill acquisition (beginner, novice, competent, proficient, expert) and the differences between them, again being overly generic, with no premise in surgical training, while also preceding the incorporation of CBME by decades (Dreyfus, 1980).

Secondly, delineating the methods by which technical competence is assessed and their inherent benefits and drawbacks can help educators and program directors determine which methods they can incorporate into their training systems (based on their current needs, resources and logistics). Demonstrating that these assessment methods for the most part are valid and reliable provides at least some credibility for their use by the various General Surgery training programs and governing bodies (Szasz et al., 2015). In this regard, although this review was undertaken in late 2013 and completed in 2014, the majority of studies included still utilized the traditional and archaic concepts of validity and reliability (and therefore these concepts were used in the review), not Messick's conceptual framework (Messick, 1995; Szasz et al., 2015). Although this conceptual framework has been around for quite some time, it's uptake in

education and surgical education in particular has been slow, with many educators for years calling for the discontinuation of the old and full incorporation of this new framework (D. A. Cook et al., 2014; Korndorffer et al., 2010). Our review had the opportunity to once again bring to light this slow incorporation of Messick's framework in surgical education and reignite this conversation (Szasz et al., 2015). A review also published in the Annals of Surgery around the same time as ours (2014-2015) evaluated various technical performance instruments and whether the included studies met the five criteria of Messick's framework, as determined by the authors, not necessarily the included studies (Ghaderi et al., 2015). The timing of our review occurred at the precipice of this discussion in surgery and when combined with the review by Ghaderi *et al*., (although it has many shortcomings), appears to have made the surgical education community more aware about incorporating this framework (Ghaderi et al., 2015; Szasz et al., 2015; Todsen & Ringsted, 2015). This is evidenced by recent studies utilizing this conceptual framework in Orthopedic Surgery, Obstetrics & Gynecology, Urology and General Surgery (MacEwan, Dudek, Wood, & Gofton, 2016; Thinggaard et al., 2015).

Thirdly, delineating the methods by which technical competence is assessed and demonstrating that there are roughly 85 such ways, lends credence to recent views that there are too many assessment instruments currently used in medical/surgical education (Ghaderi et al., 2015; Jelovsek et al., 2013; Szasz et al., 2015). Even with the 85 included studies, our review omitted the methods by which technical skill, ability and proficiency are assessed; if included, the total number of assessment instruments would number in the hundreds (Szasz et al., 2015). How can training programs and governing bodies keep track of the many instruments utilized and keep up to date with the new literature coming out at a rapid rate? This review in concert with the inundation of assessment instruments, suggests that perhaps educators should move away from persistently creating new instruments, and rather focus on how to use the results of the assessment instruments that are already in existence. By collecting evidence to support their use in line with Messick's conceptual framework of validity, and by setting performance standards to differentiate between scores on such assessments (S. Messick, 1989; Messick, 1995; J. J. Norcini, 2003).

In summary, our systematic review has many novel attributes that can help operationalize CBME in General Surgery (Szasz et al., 2015). The first step in assessment is ensuring that everyone speaks the same language; this study purports a definition of competence that training

programs and governing bodies can utilize (Szasz et al., 2015). Although this definition specifically relates to technical competence, it is more granular a concept than the definitions previously created by its precursors and it can be adopted and adapted for the other competencies/domains within CBME as well (Dreyfus, 1980; J. R. Frank, Snell, et al., 2010; Satava et al., 2003). This review in concert with other work done around the same time also supports the transition of surgical education away from the traditional concept of validity towards Messick's conceptual framework, a long needed evolution (Ghaderi et al., 2015; Korndorffer et al., 2010). This move towards the conceptual framework of validity ensures that everyone is on the same page regarding psychometric properties and helps further support the use of certain assessment instruments within CBME, with the ongoing collection of evidence for their use. Finally, this review supports recent opinions that instead of creating additional assessment instruments, the education community should focus on what the results of such assessments mean (Ghaderi et al., 2015; Jelovsek et al., 2013). Only then can a select few instruments (those with accumulating evidence) be incorporated into CBME, in concert with performance standards, as described below.

7.2.2 International assessment practices

With the ongoing transition towards CBME by various jurisdictions, and this occurring at differing rates, with some (UK) more advanced than others (Canada and the United States) in relation to the implementation of assessment modalities, there are likely many actual assessment practices ongoing for which little literature is available ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014; "Good Medical Practice," 2014; Holmboe et al., 2015; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). The international education directorates' study was therefore completed in addition to the systematic review described above. Given that CBME is a paradigm shift across all levels of medical/surgical training we sought performance assessment information across all such phases: selection into training, in training progression, and certification (C. Carraccio et al., 2015).

The implications of this study's findings in relation to CBME are twofold. Firstly, it provides information on current and ideal technical assessment practices, thereby sharing cross-jurisdictional best practice protocols in the anticipation of implementing such assessments into

CBME. Secondly, it provides further real world confirmation on where such performance assessments should take place (the workplace).

This international directorates study describes that the current and ideal performance assessments don't differ all that much (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). Currently for in-training and certification time points, the majority of international jurisdictions utilize direct observations of trainees utilizing either a global rating or task-specific instrument (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). In an ideal setting these same instruments would be utilized, along with the use of trainee ITERs and less so other modalities, including case-logs (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). Issues with ITERs and other such modalities are described in the introduction of this thesis, section 1.4.2, 1.8.5 and the specific discussion section of this study, but briefly, they focus on their imprecision especially for trainees that may be struggling as well as the reliance on self-reported data that may not be representative of actual competence or the use of surrogate markers that are used to infer competence (L. S. Feldman et al., 2004; Ginsburg et al., 2013; Lonergan et al., 2010; Patel, 2015; Szasz et al., 2015). ITERs although utilized as one of the main modalities for summative assessments in Canada, the United States and the UK, cannot form the main component of competency-based assessments (Compeau et al., 2009; Eardley et al., 2013; Meyerson et al., 2014). Along the same line, neither can self-reported or surrogate outcomes. The reasons for this center on the basic tenants that at least theoretically underpin competency-based assessments for technical performance (although this also applies to non-technical performance). Firstly, such assessments should as much as possible occur in the workplace whether that's the OR, ward or clinic (Holmboe et al., 2010; Potts, 2016; Rekman, Gofton, Dudek, Gofton, & Hamstra, 2016). Secondly, although somewhat intuitive, these assessment instruments must have the capacity to be directly used in the workplace for each encounter (Holmboe et al., 2010; Rekman et al., 2016). Thirdly, and as discussed above, evidence must be collected to support the interpretation of the results created by these instruments and standard setting should be utilized to differentiate appropriate levels of performance (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; Ghaderi et al., 2015; Messick, 1995). Finally, skilled and trained raters must be utilized (Holmboe et al., 2010; Pugh et al., 2015). Only if all of these tenants are met, will such assessments be appropriate for CBME. ITERs and case logs, do not meet any of these prerequisites. These four

tenants should underpin assessment in CBME whether that assessment is formative or summative in nature. There are specific attributes within each type of assessment that are also required and these are discussed below in section 7.4

The issue of the accuracy of faculty observations is a recurring theme in this thesis. In our international directorates study, it was demonstrated that faculty undertook the majority of assessments, using either a global or task specific instrument as described above (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). At the time of this manuscripts publication our research group supposed that faculty assessments were inferior to assessments performed by external individuals and we argued for the use of 'objective' external assessments (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016; Szasz et al., 2015). Subsequently, based on these views, this was seen as a weakness in this study's current and ideal assessment practices cross jurisdictionally. Our impression that faculty assessors are inferior was for the most part predicated on anecdotal evidence and incomplete and outdated research (Elliot & Hickam, 1987; Kalet et al., 1992; Wanzel, Ward, et al., 2002). Yet no studies directly comparing faculty and external raters in the OR had been completed. Therefore, this premise was directly evaluated in our study described in Chapter 5 and discussed below in section 7.4. The international study demonstrating that the majority of jurisdictions utilized faculty assessors is no longer seen as a weakness, but rather an opportunity for the implementation of competency-based assessments using available and logistically practical resources.

The international directorates study also confirmed cross-jurisdictionally in a real world setting, what has been stated by the theoretical competency-based assessment literature (Holmboe et al., 2010; M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). That assessment of technical performance should occur in the OR (workplace) rather than the simulation laboratory (Crossley & Jolly, 2012; Holmboe et al., 2010). Although there are some benefits to the simulated setting mostly focusing on procedural reproducibility, the work by Miller *et al*., and others, has documented a discordance between how trainees perform in a constructed setting, compared to how they perform in the real world (Crossley & Jolly, 2012; Holmboe et al., 2010; Miller, 1990; Rethans, Sturmans, Drop, van der Vleuten, & Hobus, 1991). Given that patients are operated on in the OR, assessment of performance should also occur

there. The fact a trainee can perform a task or procedure in a simulated setting is neither informative to the training program nor reassuring to the public.

In summary, our international directorates study has many novel attributes that can help operationalize CBME in General Surgery (M. Louridas, Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G, 2016). It demonstrates current and ideal technical performance assessment practices. This sharing of cross-jurisdictional protocols can be utilized in the planning and implementation phases of competency-based assessments as the transition towards CBME occurs. The study also documents that current practices do not differ all that much from ideal practices based on the expert views of these education directorates – a positive finding given the likely hesitancy these jurisdictions feel about a full transition to CBME. The support for direct performance assessments and less of a focus on ITER assessments and the use of surrogate markers by these jurisdictions, provides support to the theoretical components and requirements of competency-based assessments (Holmboe et al., 2010; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016). Finally, the education directorates' real world data supports the available literature suggesting that performance assessments within CBME should occur in the OR (workplace) (Crossley & Jolly, 2012; Holmboe et al., 2010; Miller, 1990; Rethans et al., 1991).

7.3 Consensus-based training and assessment model

Over the last few decades there has been a prevailing opinion that graduating trainees are less prepared to undertake independent practice than their predecessors (Soper & DaRosa, 2014). The reasons stated point directly to how recent changes including: public opinion, work hour restrictions, fiscal constraints to training residents, the ageing population and the focus on surgeon-volume outcome data, have negatively impacted resident training (Kempenich et al., 2015; Mattar et al., 2013). These altered training experiences for residents combined with the information garnered from our two initial studies – led us to identify that a major challenge to incorporating technical competence assessments (while also being relevant for many other competencies/domains within CBME) and performance standards, stems from a lack of knowledge as to which procedures and tasks still form the scope of practice for General Surgery trainees and which specific procedures should be utilized for milestone assessments.

The implications of this study's findings for CBME center around these two topics: the creation of a new more progressive training framework specific to junior and senior level

trainees and the delineation of procedures for milestone assessments to determine whether trainees can progress in their training (Szasz et al., 2016).

The training framework presented in this study is really an evolved version of those currently utilized by training programs in Canada and the United States ("Curriculum outline for General Surgery Residency," 2015-2016; "The Future of General Surgery: Evolving to meet a changing practice," 2014; "General Surgery: Content Outline for the ABS In-Training Examination," 2013; "Objectives of Training in the Specialty of General Surgery," 2010; Szasz et al., 2016). The increased emphasis in this study on commonly performed procedures and tasks and decreased focus on subspecialized and complex procedures illustrates that the program directors are acutely aware of how the traditional training frameworks are no longer meeting the needs of the trainees and training programs, especially as the transition to CBME occurs (Szasz et al., 2016). This new training framework places more of an emphasis on optimizing resident training by increasing their competence and confidence on procedures that are commonly performed by general surgeons in practice (M. F. Brennan & Debas, 2004; Coleman et al., 2013; Fonseca et al., 2014; Szasz et al., 2016). While leaving procedures that are inadequately learned during General Surgery residency and not performed by general surgeons in practice, to fellowship trained surgeons (Al-Qurayshi et al., 2016; M. F. Brennan & Debas, 2004; Jeong et al., 2014; Ricci et al., 2014). Accordingly, what is perhaps most relevant in this study, are not the procedures and task found in the new training framework presented here, but rather those that have traditionally been represented in previous and current training guidelines, but are absent in this study (Szasz et al., 2016). This transition to a more streamlined training framework is supported by research completed by Chung *et al*., and more recently Bell *et al*. and Drake *et al*., documenting that trainees during residency complete only a small subset (and even smaller subset with appropriate case volumes) of expected procedures in the current training guidelines (R. H. Bell, Jr. et al., 2009; R. S. Chung, 2005; Drake et al., 2013). Given the prolonged focus on the outdated and extensive training guidelines, has at least partially contributed to residents completing training and being ill prepared for independent practice (Mattar et al., 2013).

The assessment framework presented here, for the first time proposes procedures for milestone assessments, the currency of CBME (Holmboe et al., 2010; Iobst et al., 2010; Szasz et al., 2016). As discussed in the introduction of this thesis, although milestones have been at the forefront of the CBME discussion, in Canada they are generically defined for all specialties, with

no specialty specific elements, which preclude their implementation into residency (J. R. Frank, Snell, L.S., Sherbino, J, 2014). Although the United States is a littler further ahead in this regard, having provided milestones for each discipline, these again are not specific enough in order to be actionable (Holmboe et al., 2015; Swing et al., 2013). The assessment framework formulated in this study provides tangible information in the form of four procedures, based on the experience and expertise of General Surgery program directors (Szasz et al., 2016). These four procedures are common and performed often, and they can now be utilized to assess trainee technical performance ("The Future of General Surgery: Evolving to meet a changing practice," 2014). Although the milestones discussed here are intended as technical milestones, it is likely appropriate and feasible to assess many of the other competencies/domains within CBME (i.e. communication, professionalism and leadership for example) during these milestone assessments as well. This collective assessment approach is supported by leaders in the field including van der Vleuten *et al*., Carraccio *et al*., and Hawkins *et al*., who purport the integration of multiple competence assessments rather than viewing each independently (C. L. Carraccio & Englander, 2013; Hawkins et al., 2015; van der Vleuten & Schuwirth, 2005). Thus, our research group with the support of the literature envision these procedures as milestone opportunities where both technical and non-technical performance can be assessed. The information gathered from these milestones assessments collectively, in coordination with how these trainees fare on more traditional evaluations that test their knowledge and judgment, can then determine whether a trainee has met specific qualifications and can progress in their training (i.e. from a more junior to senior level trainee). If the holistic performance of trainees is found to be inadequate, a remedial program directed at the specific insufficiencies can be created (Bhatti, Ahmed, Stewart, Miller, & Choi, 2015; Epstein, 2007; Wu et al., 2010). Given that these milestone assessments would be used to evaluate what trainees have learned and to make decisions about those trainees, they would be deemed summative in nature (Hawkins et al., 2015).

In summary, our training and assessment framework study has many novel attributes that can help operationalize CBME in General Surgery (Szasz et al., 2016). Similarly to developing a common language for competence within CBME as was described by our study in Chapter 3, the training component of this framework also takes a step back and lays the groundwork for the content of a new CBME curriculum for technical performance (Szasz et al., 2016). This training framework takes all of the recent changes that have negatively impacted surgical training into

consideration, providing streamlined content that General Surgery trainees need to achieve competence and confidence in prior to independent practice, thereby optimizing their training (Coleman et al., 2013; Fonseca et al., 2014; Kempenich et al., 2015; Mattar et al., 2013). Complex and specialized procedures, which are still present in the current frameworks utilized, are omitted on the behest of the program directors, in keeping with their transition to subspecialized centers completed by surgeons with fellowship training (Birkmeyer et al., 2013; Birkmeyer et al., 2002; "Curriculum outline for General Surgery Residency," 2015-2016; "The Future of General Surgery: Evolving to meet a changing practice," 2014; "General Surgery: Content Outline for the ABS In-Training Examination," 2013; Mattar et al., 2013; "Objectives of Training in the Specialty of General Surgery," 2010). The assessment framework presented here, for the first time provides actionable milestones (opportunities for summative assessments) that can be used to assess technical performance (Szasz et al., 2016). In keeping with recent literature, at the time of these milestone assessments, other competencies/domains that together make up CBME can also be assessed (C. L. Carraccio & Englander, 2013; Hawkins et al., 2015; van der Vleuten & Schuwirth, 2005).

7.4 Performance assessors

Although there are many requirements for competency-based assessments as described above in section 7.2.2, there are no documented requirements as to whom these assessors actually ought to be (their specific characteristics) (Holmboe et al., 2010; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Swing et al., 2009). While some studies have evaluated faculty (internal) assessments and compared them to more ‘objective’ measures of trainee performance or other internal assessors, none have compared the assessment results between different groups of raters in the OR, both of whom have been trained (removing lack of training as a confounding variable) (Elfenbein et al., 2015; Farrell et al., 2010; L. S. Feldman et al., 2004; Goldstein et al., 2014; Herrera-Almario et al., 2016; Moonen-van Loon et al., 2015; Ray et al., 2016; Reid et al., 2014; Steigerwald et al., 2015; Ward et al., 2003). This lack of evidence comparing and contrasting internal and external raters for use during summative and formative performance assessments and the education community’s anecdotal and preconceived notions of internal rater inferiority, led to this study (Elliot & Hickam, 1987; Kalet et al., 1992; Wanzel, Ward, et al., 2002). Given the use of internal raters comes with many logistical and financial advantages, starting the conversation in this regard as this study has, and subsequently collecting further

convincing and robust evidence is imperative towards operationalizing assessment in CBME (Chen, 2015a; Takahashi, 2011).

The implications of this study's findings for CBME differ depending on whether we're talking about their use for formative or summative assessments. Nonetheless, documenting that a moderate to strong association exists between internal and external assessors for technical and non-technical performance scores, respectively, after training, lends credibility for the use of such internal assessors in surgical education (M. Feldman et al., 2012).

Although the purpose of our study was to delineate whether an association exists and if so the degree of that association between internal and external raters and subsequently to determine overall differences (stringency) – we did not base the study in either a summative or formative framework. In fact, we wanted to collect evidence and compare our findings with other available literature to determine whether our results can support either use. In terms of formative assessments, the focus was on what facets are necessary for their incorporation into surgical training and how our study does or does not meet such facets. In terms of summative assessments, the focus was on how accurate is accurate enough, when comparing the use of different rater types.

The required guidelines to implementing formative assessments, in addition to what is required for competency-based assessments include: 1) the performance assessment instrument should match or at least closely resemble the instrument used for summative assessments, 2) formative performance assessments should evaluate the same domains that would be evaluated for summative assessments, 3) assessments should occur in the workplace, 4) trainees undergoing formative assessments should be aware of what is expected of them, 5) the assessments should occur immediately after an observed performance, and 6) feedback should occur at the completion of each formative assessment and be specific in regards to the task that was completed (Allal, 2005; Rolfe & McPherson, 1995; William, 2011). Our study findings meet all of these requirements except for the portion regarding feedback. However, as can be deduced from both the moderate to strong association between external and internal raters as well as the information gathered by the research group during rater training (that the internal raters were able to identify, transcribe and verbalize the full spectrum of behavioral exemplars of technical and non-technical performance). Combined with the fact that these internal raters are all staff

surgeons with experience in teaching trainees and an understanding of trainee education guidelines, they very likely should be able to provide appropriate feedback targeted at trainee improvement ("The Future of General Surgery: Evolving to meet a changing practice," 2014; "Objectives of Training in the Specialty of General Surgery," 2010; Ramani & Krackov, 2012). While the incorporation and implementation of trainee feedback was not the focus of this study. The results presented here, with our adherence to the features required for formative assessment implementation (described above and in the introduction, section 1.5.4), and the adherence to the components within competency-based assessments described above, provide credible evidence for the use of internal rater formative assessments in CBME (Holmboe et al., 2010; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Swing et al., 2009).

As stated by Rolfe *et al*., the only differentiation between formative and summative evaluations is really the intent of the assessment (Rolfe & McPherson, 1995). In that regard summative assessments follow all of the components of competency-based assessments described above and share some of the features of formative assessments (Allal, 2005; Holmboe et al., 2010; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Rolfe & McPherson, 1995; Swing et al., 2009; William, 2011). In addition to this, given that summative assessments lead to consequences on the part of the trainees and training programs, there is an increased focus on: 1) the psychometric properties of the assessment instruments (even if these are the same instruments that were utilized for formative assessments), and 2) the use of appropriate pass/fail criteria that have been developed in a prescribed and systematic approach (Cizek & Bunch, 2007c; Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010; R.J. Mislevy, 2003; R.J. Mislevy, 2011; Ten Cate et al., 2016). Although the latter point is not directly addressed in our Chapter 5 study, it forms the basis of our study described in Chapter 6 of this thesis and it is discussed in full detail, below in section 7.5. The former point is directly addressed in the study from Chapter 5, as it was our goal to merge the work from that study with what is already known from using both the OSATS (J. A. Martin et al., 1997) and OSANTS (Dedy, Szasz, et al., 2015) rating instruments. This goal is in keeping with the continual nature of the validation process as stated by Messick in 1989 “validity is an evolving property and validation is a continuous process” (S. Messick, 1989). Both the OSATS and OSANTS assessment instruments have been shown to meet many of the traditional facets of validity and reliability, including construct validity, internal consistency and inter-rater reliability (Dedy, Szasz, et al., 2015; J. A. Martin et

al., 1997; Reznick et al., 1997). Furthermore, they meet many of the criteria within Messick's conceptual validity framework as discussed in the introduction, section 1.5.3 (Dedy, Szasz, et al., 2015; S. M. Downing, 2003; Ghaderi et al., 2015; J. A. Martin et al., 1997; Messick, 1995; Reznick et al., 1997). This particular study described in Chapter 5 adds evidence to support the interpretation of the results produced by both of these assessment instruments mostly in the categories of content evidence and response process – by ensuring that the internal raters were experts in General Surgery, had an understanding in trainee education and assessment and were trained (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995).

Using assessment instruments, which meet many of the criteria within Messick's conceptual framework, the findings of this study provide initial evidence where previously there was none, regarding the assessment accuracy of internal assessors (i.e. how accurate is accurate enough) (S. Messick, 1989; Messick, 1995). The major differentiation between external and internal assessors is independence – whereas external assessors have no relationship with the program or trainee under assessment, internal assessors have some underlying relationship and therefore may have a stake in the assessment outcome (Chen, 2015a). As discussed in section 7.2 above, for many years it was presumed that individuals who may have a stake in the assessment outcome, are inferior for use during performance assessments (particularly those that are summative) (Elliot & Hickam, 1987; Kalet et al., 1992; Wanzel, Ward, et al., 2002). Furthermore, as discussed in the introduction section 1.8, specific features of the faculty/trainee relationship have also been cited to preclude using such internal assessors during assessments, perhaps including those within CBME (Reid et al., 2014; T. J. Wilkinson & Wade, 2007; Williams et al., 2003). Conversely, the use of such internal assessors can come with some benefits – mostly directed at the cost and logistics of their undertaking (Chen, 2015a; Takahashi, 2011). Despite the importance of determining whether such raters can or cannot be utilized for assessments and the significance this will play in CBME, no research directly evaluating this concept has been completed. Although information can somewhat be gathered from other studies, these studies make comparisons between either internal assessors and other more formal assessments or various types of internal assessors (i.e. staff surgeons/physicians and residents) (Elfenbein et al., 2015; Farrell et al., 2010; L. S. Feldman et al., 2004; Goldstein et al., 2014; Herrera-Almario et al., 2016; Moonen-van Loon et al., 2015; Ray et al., 2016; Reid et al., 2014;

Steigerwald et al., 2015; Ward et al., 2003). The findings from our study for the first time provide evidence as to the accuracy of internal rater assessments in the real OR for both technical and non-technical performance, by comparing them to an external expert rater. Suggesting that internal raters despite a lack of independence and even a stake in the outcomes of evaluations, once trained can provide accurate trainee assessments for both technical and non-technical performance. Although a promising start, further research in other disciplines is needed to corroborate these findings and more definitively determine whether internal assessors can be utilized for summative assessment.

In summary, this study has many novel attributes that can help operationalize CBME in General Surgery. The identification in this study that internal raters correlated well with their external rater counterparts, albeit providing less stringent results, for the first time provides credible evidence for the use of internal raters for formative assessments. Although this study adhered to many of the facets required for the implementation of summative assessments, and for the first time documented that despite being tied to the training of residents, internal raters are accurate assessors of trainee performance, once trained. Further research is needed in other surgical and/or medical disciplines providing a more robust collection of evidence for utilizing internal assessors during summative assessments.

7.5 Performance standards

This thesis in attempting to help operationalize CBME culminated in the creation of technical and non-technical performance standards in General Surgery that can be utilized for summative assessments. There is currently a paucity of such standards in surgical education. In their absence, the scores on technical and non-technical performance assessments during training, even when following all of the prescribed guidelines required for competency-based and summative assessments are arbitrary, providing no real information whether a trainee has attained an appropriate level of performance (Cizek & Bunch, 2007c; Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010; R.J. Mislevy, 2003; R.J. Mislevy, 2011; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Swing et al., 2009; Ten Cate et al., 2016). The ability of such summative assessments to delineate whether trainees have attained a particular standard however, is at the center of CBME and without such standards; summative assessments within CBME cannot be implemented (Cizek & Bunch, 2007c; Hawkins et al., 2015; Holmboe et al.,

2010; Koehler & Nicandri, 2013). Furthermore, a lack of evidence describing the relationship between trainees who are technically competent and non-technically competent, or vice versa needs further investigation ("Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0," 2012; Hull et al., 2012). Finally, as the transition to CBME continues, there is a need to collect evidence comparing the traditional and CBME systems in order to garner support for this transition. Therefore, identifying predictive factors for performance standard acquisition and relating these back to the traditional and CBME frameworks is essential - providing validity evidence of sorts to this transition.

The implications of this study's findings in relation to CBME relate to providing specific evidence directed at these three identified voids. Firstly, this study demonstrates that utilizing the contrasting groups methodology, standards for technical and non-technical performance could be set. These performance standards were also found to be credible and reliable. Secondly, this study documents a disparity between trainees who are technically and non-technically competent, information that is important as educators continue to devise and incorporate CBME training curricular, focusing on all seven competencies/domains (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015). Thirdly, as the transition to CBME occurs, collecting evidence comparing the traditional and the CBME training frameworks is important, if only to ensure educators are not missing the mark when it comes to CBME, helping persuade naysayers to buy in to the transition and the fact that for at least the foreseeable future a hybrid model will likely coexist (C. Carraccio et al., 2002; C. L. Carraccio & Englander, 2013; Iobst, 2015; Leung, 2002).

The performance standards created in this study fill a major void in the implementation of competency-based assessments into surgical training (Cizek & Bunch, 2007c; Koehler & Nicandri, 2013). As reviewed in the introduction and discussion sections of this particular study, milestones are opportunities for summative assessments (Cizek & Bunch, 2007c; Cogbill, 2014; J. R. Frank, Snell, L.S., Sherbino, J, 2014; Holmboe et al., 2010; T. J. Wilkinson et al., 2001). In order for these milestone assessments to become functional, there must be a clear delineation between trainees who have met the specific milestones (are competent) and those that have not. This is accomplished by setting performance standards, which then provide defensible evidence to all involved stakeholders (including the trainees themselves) whether an appropriate level of performance has been reached at the milestone assessment (S. M. Downing, Tekian, A,

Yudkowsky, R, 2006; J. R. Frank, Snell, L.S., Sherbino, J, 2014). In turn these milestones can then serve as the blueprint for the creation of summative assessments (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Green et al., 2009; Hawkins et al., 2015; Holmboe et al., 2010). As supported by evidence, examinee-centered strategies such as the contrasting groups methodology, have been demonstrated to be a valuable modality for standard setting during clinical encounters (Cizek & Bunch, 2007d; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). Although the contrasting groups methodology specifically has been utilized in other areas of medicine and even surgery, for the most part these studies were in the simulated setting, none of the ones that occurred in the workplace involved General Surgery trainees and no studies assessed non-technical performance (J. D. Beard et al., 2005; S. de Montbrun, Satterthwaite, et al., 2016; Diwadkar et al., 2009; Jacobsen et al., 2015; Konge et al., 2013; Konge et al., 2012; Preisler et al., 2015; Sedlack, 2011; Sedlack et al., 2016; Thinggaard et al., 2015; Tolsgaard et al., 2014).

For the first time, our work set standards for the laparoscopic cholecystectomy for General Surgery trainees using a group of appropriate expert judges, directly in the workplace setting, meeting all of the components required of competency-based assessments and summative assessments (which necessitate standard setting) (Cizek & Bunch, 2007b, 2007c; Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010; Livingtson, 1982; J. J. Norcini, Shea, JA, 1997; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Swing et al., 2009; Ten Cate et al., 2016). These standards focus on both technical and non-technical performance, the latter of which has never been completed before, yet is imperative as CBME focuses on all seven competencies/domains, not just medical expertise and technical performance (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015). The benefit of utilizing the contrasting groups methodology is the ability to adjust the standard, depending on the purpose of the assessment, in order to minimize the most detrimental consequence of making a false decision for that particular assessment (either false positive or false negative) (S. M. Downing, Tekian, A, Yudkowsky, R, 2006; J. J. Norcini, 2003). As reviewed in the discussion section of this specific study, if the summative assessment is completed during training and there will be subsequent summative opportunities for trainees to demonstrate their abilities, minimizing the false negative decisions, as these can be costly to both the training program and trainee is important. If however, the summative assessment occurs at the completion of training (certification) and

determines whether the trainee can undertake independent practice, minimizing the false positive decisions is imperative for both the trainee and the public. Within our study, as the purpose of the assessment is not yet delineated, false positive and false negative decisions were seen as equally detrimental and therefore the performance standard was not adjusted.

We also outline that our performance standards are credible by basing them on the three components within Norcini's framework namely: the creation of realistic standards, utilizing criterion referenced methodologies (contrasting groups) and using multiple, varied expert judges (J. J. Norcini, Shea, JA, 1997). Norcini's credibility framework is similar to Messick's conceptual framework of validity only now for standards (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995; J. J. Norcini, Shea, JA, 1997). In essence, the adherence to Messick's framework provides evidence to support the results (scores) garnered from specific assessment instruments, while the adherence to Norcini's framework provides evidence to support the interpretation of the performance standards, created using these results (scores) (D. A. Cook & Beckman, 2006; S. M. Downing, 2003; Ghaderi et al., 2015; Messick, 1995; J. J. Norcini, Shea, JA, 1997). Therefore, both are imperative to creating appropriate standards. Our study also demonstrates excellent performance standard reliability when comparing competent/non-competent decisions based on expert judgment and performance standard attainment. Taken together, this focus on credibility (validity) and a high level of reliability for our performance standards contributes to the psychometric properties required for summative assessments as described in section 7.4 (Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010; Ten Cate et al., 2016). Finally, as the study evaluating internal assessors discussed in section 7.4 contributed content and response process evidence in regards to Messicks' framework towards the use of OSATS and OSANTS assessment instruments, this study provides evidence in the remaining three categories including internal structure, relationship to other variables and consequences (D. A. Cook & Beckman, 2006; Dedy, Szasz, et al., 2015; S. M. Downing, 2003; Ghaderi et al., 2015; J. A. Martin et al., 1997; Messick, 1995).

The demonstration in this study that there was a 21% disparity between trainees who were found to be technically and non-technically competent is quite interesting. Non-technical performance has entered the dialogue of medical education over the last fifteen years or so (R. Flin, Martin, R., Goeters, K.M., Hormann, H.J., Amalberti, R., Valot, C., Nijhuis, H, 2003; R. Flin, O'Connor, P, Crichton, M, 2008; Youngson & Flin, 2010; Yule, Flin, Paterson-Brown, &

Maran, 2006). Since that time, there have been a number of studies completed, demonstrating that non-technical failures on the part of surgeons or trainees contribute to negative patient outcomes (Boet et al., 2014; Firth-Cozens & Mowbray, 2001; Greenberg et al., 2007; Lingard et al., 2004; Mishra et al., 2008). There is however, a limited amount of literature assessing the influence that non-technical performance has on technical performance, save for a review compiled by Hull *et al*., documenting mixed results, for the most part evaluating trainees not in the workplace using technical and non-technical assessment instruments not adhering to Messick's framework of validity (Hull et al., 2012; S. Messick, 1989; Messick, 1995). More recently, studies in Anesthesiology have been completed in a simulated setting using appropriate assessment instruments comparing technical and non-technical performance, again documenting mixed results (Gjeraa, Jepsen, Rewers, Ostergaard, & Dieckmann, 2016; Phitayakorn et al., 2015; Riem, Boet, Bould, Tavares, & Naik, 2012). Furthermore, in addition to comparing these performances, until the completion of our study, there was no data available evaluating the ability of trainees to meet a technical and non-technical performance standard concurrently during the same operative procedure. Given the disparity demonstrated in this study, may go against the traditional views that as trainees progress through training, learning and demonstrating their technical performance abilities, they also acquire non-technical abilities ("Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0," 2012; Greig et al., 2015). This data combined with a limited amount of non-technical teaching in surgical programs that are not for research purposes, provides an avenue for change within the CBME framework (Gjeraa et al., 2016; Greig et al., 2015; Nicksa, Anderson, Fidler, & Stewart, 2015; Phitayakorn et al., 2015; Riem et al., 2012). This change could be in the form of explicit non-technical skills training and subsequent assessment that is both formative and summative in nature, the latter of which will require performance standards to determine whether an appropriate level has been reached - as were set in this study. This incorporation of non-technical skills training will enable the many other competencies/domains required of trainees to be taught and assessed as the transition to CBME occurs (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015).

Finally, this study identifies predictive factors for acquiring both the technical and non-technical performance standards. This provides an opportunity to compare and contrast the traditional and CBME frameworks. The traditional framework, based on time used PGY level for

the most part to determine when trainees had acquired all of the attributes required for independent practice (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Hodges, 2010). While the CBME framework, based on competencies predicates itself on using observable performances to determine when trainees have acquired the attributes required for independent practice (J. R. Frank, Snell, L.S., Sherbino, J, 2014; Hodges, 2010). As trainees move through this CBME framework there will have to be a way to determine how many formative assessments are required and when a trainee is ready for their summative assessment, and this can be accomplished by evaluating trainee case numbers. We are not by any means suggesting that case numbers should be used as a surrogate of competence as this is accompanied by many issues, discussed above in sections 1.6, 3.5 and 7.2, but rather that case numbers can be used to provide the trainee and training program and idea of when a resident should be able to approach competence (Lonergan et al., 2010; Szasz et al., 2015). These case numbers can also be compared to the PGY level in the traditional framework, looking for parallels or where the trainee would have been in that PGY level in reaching the performance standards. The eventual education goals of many governing bodies are to fully transition towards a CBME system ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014; "Good Medical Practice," 2014; Holmboe et al., 2015; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). Currently and for the foreseeable future however, time based training will continue to play a role, likely in a hybrid model as a full transition to CBME occurs overs years or even decades (Iobst, 2015). Therefore this is an opportune time to compare the two and ensure that educators are not missing the mark as it comes to the implementation of CBME. The demonstration in this study that three and a half years of training and number of procedures performed (46.5 for technical and 35 for non-technical) are both equally able to predict acquiring the performance standards, lends credibility to this comparison. These findings are also supported by research previously carried out, where graduating trainees' case numbers for the laparoscopic cholecystectomy, when extrapolated out would provide a case number in our range (35 to 46.5) somewhere between their third and fourth year of training in a traditional training framework (R. H. Bell, Jr. et al., 2009; Carlsen et al., 2014; R. S. Chung, 2005). This information provides supporting evidence for the transition towards a CBME framework.

In summary, this study has many novel attributes that can help operationalize CBME in General Surgery. The demonstration that credible and reliable performance standards can be set in General Surgery while adhering to the requirements of competency-based and summative assessments, provides evidence for their use during milestone assessments (Cizek & Bunch, 2007b, 2007c; Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010; Livingtson, 1982; J. J. Norcini, Shea, JA, 1997; Potts, 2016; Pugh et al., 2015; Rekman et al., 2016; Swing et al., 2009; Ten Cate et al., 2016). The information from this study regarding a disparity in concurrent technical and non-technical standard acquisition, provides the first piece of evidence in this regard, and supports the use of explicit non-technical skills training and assessment that will focus on the other competencies/domains within CBME (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015). While the predictive factors in this study demonstrate to educators that they are not missing the mark in their transition away from traditional training, towards CBME. Allowing case numbers to suggest approximately when trainees are expected to reach competence for a particular procedure, both technically and non-technically (and thus when they may ready for a summative assessment).

7.6 Conclusions

The studies contained within this thesis collectively provide evidence that can support the transition to CBME and assessment in Canadian General Surgery training. They focus on an education and assessment model that is in tune with recent changes affecting surgical training and provide granular operative training guidelines and specific procedures for milestone assessments. The instruments that should be used to assess technical performance are depicted and the essential properties of these assessment instruments are described - including the need to adhere to the conceptual framework of validity in order to collect evidence to support their use. An international perspective delineates the sharing of cross-jurisdictional best practice protocols that can be useful in planning and implementing CBME and assessment, while also demonstrating that current and ideal assessment practices do not differ all that much, and providing real world support for the theoretical components of CBME and assessment. The final two studies in this thesis focus on operationalizing technical and non-technical performance assessments into General Surgical training. These studies demonstrate that trained internal raters are accurate, albeit less stringent assessors than their external rater counterparts, for both technical and non-technical performance, likely appropriate for formative type assessments, with

further work in other areas of medicine and surgery required to determine whether they can be used for summative assessments. The thesis culminates by setting credible and reliable performance standards for both technical and non-technical performance, while identifying an inability by some trainees to concurrently achieve both standards and demonstrating factors that can predict standard acquisition. These standards can be utilized for summative type assessments (milestones) that need to be achieved for trainees to progress in their training, while the information on concurrent standard acquisition and predictive factors, can be utilized to structure training curricula within CBME and to ensure the transition to CBME isn't occurring with too much haste, respecting and regarding the traditional training framework.

CHAPTER 8

GENERAL LIMITATIONS

Chapter purpose

This chapter focuses on the general limitations and mitigating factors (where applicable) of the overall thesis. These limitations include 1) institutional and faculty level differences that may have impacted trainee performance and were unable to be captured, 2) trainee participation in other educational research endeavors which may have confounded their performance in our research studies, and 3) a deficiency in formally assessing stakeholder buy in for the implementation of our research findings. These general limitations are in addition to the study specific limitations contained in the discussion section of each chapter.

8 General Limitations

8.1 Institutional and faculty level differences

In the studies described in Chapter 5 and 6, there are multiple factors at the institutional and faculty level, which were not captured and may have affected (in either a positive or negative manner) the performance of the trainees. Although not captured, the goal of this research was to assess trainees in the workplace setting under 'real' conditions (not standardized or ideal conditions), as this is the basis of CBME, regardless of trainees' previous experience (Hawkins et al., 2015; Holmboe et al., 2010). Furthermore, the goal of our study in Chapter 6, was to set authentic performance standards, using the contrasting groups methodology, with the initial division between competent and non-competent trainees based on actual performance. Not an artificial distinction such as novice versus expert or total number of procedures performed as has been previously done (Jelovsek et al., 2010; Konge et al., 2012; Preisler et al., 2015; Thomsen et al., 2015).

At the University of Toronto, General Surgery trainees rotate through various academic hospitals during their residency. Although the main foundations and General Surgery curriculums are geared at all of the residents, regardless of their hospital allocation, some hospitals focus on various aspects of surgical teaching and specific techniques more than others (laparoscopy for example), simply because they specialize in disease processes that are more amenable to those techniques (bariatric surgery, colon and rectal surgery etc.) leading to a variability in trainee experience between hospitals ("Objectives of Surgical Foundations Training," 2014; "Specialty Training Requirements in General Surgery," 2015). Furthermore, the General Surgery faculty at the various hospitals all have expertise in different clinical and research areas and subsequently teach around those areas, some of which are educational or human factors (non-technical skills) related, again creating a disparity in trainee exposure between hospitals. Consequently, the residents who rotated at St. Michael's Hospital where the data collection for our studies occurred all arrived from different hospital sites, where their previous experience or lack of experience may have contributed to their overall performance, in addition to their actual ability. This can also be said about the University of Calgary and Western University, although unlike at the University of Toronto where there are 11 affiliated hospitals, they have five and three (plus community affiliations), respectively, leading to likely less overall

institutional variability ("General Surgery - Canadian Resident Matching Service (CaRMS)," 2015).

In keeping with institutional differences on the larger scale, the foundations curriculum and the senior resident curriculum that all General Surgery residents undertake during their training regardless of University training program is outlined by the RCPSC and is therefore somewhat standardized ("Objectives of Surgical Foundations Training," 2014; "Specialty Training Requirements in General Surgery," 2015). In addition to this formal curriculum geared mostly at knowledge and judgment, certain Universities may have additional technical and non-technical skills workshops that may have impacted residence performance on the assessments within our studies, but again may not be formalized or explicit and therefore were not captured here.

At the faculty level, variability in demeanor, behavior and teaching style, which can differ drastically amongst staff, may also have impacted trainee performance (Anderson et al., 2013; Apramian, Cristancho, Watling, Ott, & Lingard, 2015; McMains, Peel, Weitzel, Der-Torossian, & Couch, 2015; N. K. Roberts, Brenner, Williams, Kim, & Dunnington, 2012). Reasons for this could include staff temperament, age, location and type of training program completed, the presence of advanced research degrees (some in education), and previous experience with trainee education and assessment. In the OR, the amount and style of teaching, the staff surgeon's comfort with the trainee, the case and the patient can all contribute in creating a situation that can encourage or discourage trainees, impacting their performance (Apramian et al., 2015; Bhandari et al., 2003; N. K. Roberts et al., 2012). Outside of the OR, pre-operative teaching around the technical and/or non-technical aspects of a case may also subsequently impact the performance of a trainee in the OR (Anderson et al., 2013).

These are inherent shortcomings that afflict most if not all research in medical/surgical education, as the trainees undergoing assessment are concurrently training in their specific specialties. These shortcomings are somewhat mitigated in studies evaluating performance in the workplace (such as ours) as the goal is not to complete standardized assessments of trainees who have undergone the exact same training regimen, but rather, real world assessments of trainees at different training time-points. As CBME becomes more integrated into residency training with the eventual use of a specific training model as discussed in Chapter 4 and a focus on both

formative and summative assessments at specific milestones, CBME training will altogether become less variable and more formalized. This will then perhaps negate practices that contribute to inconsistencies/variable training experiences, which may currently positively or even negatively impact trainee performance.

8.2 Trainee participation in other education research

The trainees who took part in the studies in Chapter 5 and 6 may have previously taken part in educational research, which could have confounded their performance.

At the University of Toronto, studies have previously been completed (between two and five years ago) in surgical education, focusing on the implementation of various teaching curricula and coaching to improve trainee performance in the operating room (Bonrath, Dedy, Gordon, & Grantcharov, 2015; Dedy, Fecso, et al., 2015; Palter & Grantcharov, 2012). These studies often employed senior residents using a randomized methodology where an intervention (either a curriculum or individualized coaching) was compared to conventional residency training, demonstrating improved operative performance amongst trainees in the intervention groups (Bonrath et al., 2015; Dedy, Fecso, et al., 2015; Palter & Grantcharov, 2012). Although participation in such research, especially for the trainees in the intervention groups may improve trainee performance in our studies and thereby confound our results, there are multiple factors that mitigate against this. Firstly, literature has documented that retention of knowledge/skill from specific interventions and their resultant effects last variable amounts of time measured in months, not years (Edelman, Mattos, & Bouwman, 2010; Rivard et al., 2015). Secondly, these research studies for the most part focused on procedures other than the laparoscopic cholecystectomy, with no current evidence to suggest there is a transfer of skills between different laparoscopic procedures. Thirdly, most if not all of the trainees who took part in the education studies above, have now graduated as data collection occurred years ago and focused on senior residents. Finally, even if some trainees are still in the training program, the fact that all research occurred using trainees rotating through St. Michael's Hospital, these trainees are very unlikely to rotate through the same site in subsequent years, so most trainees have not been back for at least two years, again removing the likelihood of being at the site when the above mentioned studies took place and also during our studies. Nonetheless even with these mitigating factors, it is possible (although unlikely) that a few senior trainees may have participated in one

of the above-mentioned research studies and the impact of this on our results is unknown. At Western University and the University of Calgary, there is a dearth of education research previously completed, thereby minimizing the risk of such research confounding trainee performance.

As was the case in section 8.1 above, the goal of our studies was to assess trainees in the real world setting at different time points. Thus, even if trainees previously took part in a research study, our results more closely reflect the outcomes that would be expected once CBME and assessment are implemented, where trainees will undoubtedly continue to take part in education and research endeavors that may impact their performance across various competency-based assessments.

8.3 Deficiency in formally assessing stakeholder buy in

There are multiple stakeholders who in part contributed to and are interested in the results of this research. These stakeholders can be divided into two main groups. The first group is composed of stakeholders directly associated with the University of Toronto, while the second group is composed of stakeholders external to the University of Toronto. This first group consists of 1) the postgraduate medical education (PGME) office, responsible for the implementation, maintenance and accreditation of all residency programs at the University of Toronto, 2) the postgraduate education committee (PGEC), responsible for the goals and objectives of training and all education and assessment curricula in General Surgery at the University of Toronto, 3) general surgeons at the University of Toronto, and 4) the General Surgery residents at the University of Toronto. The second group consists of 1) the RCPSC, as discussed in the introduction of this thesis, the overarching governing body for all Canadian medical/surgical specialties, responsible for nationwide program accreditation, trainee certification and the maintenance of certification, and 2) the College of Physicians and Surgeons of Ontario (CPSO), the authoritative body regulating the safe practice of medicine in order to protect the public, responsible for monitoring practice standards ("Royal College of Physicians and Surgeons of Canada: History," 2014; "Self-Regulation and the Practice of Medicine," 2015).

Although our research was supported and funded by the RCPSC and incorporated contributions from some of these stakeholder groups, our goal was to remain at an arms length of these groups and their individuals, as not to influence the process or outcomes of our research

and to eliminate potential disadvantages of incorporating stakeholders during the research process (Chen, 2015b; Keith, 1993). These disadvantages include: an overly influential stakeholder group pushing their own initiatives, stakeholder groups being unable to share their opinions or feeling influenced or coerced by the other groups, an inability to reach decisions, and certain stakeholders focusing too much on the implementation of a program at the expense of research rigor (Chen, 2015b; Keith, 1993). The time and cost associated with involving multiple stakeholder groups are also prohibitive disadvantages (Chen, 2015b; Keith, 1993).

While this deficiency in formally assessing stakeholder buy in was intentional on the part of our research group and supported by the RCPSC. At this current juncture, although further research into CBME and assessment is warranted and discussed thoroughly in the future work section below, incorporating stakeholder buy in for the first time seams a reasonable option. This is particularly true in order to help implement and sustain the training and assessment model discussed in Chapter 4 and the technical and non-technical performance standards discussed in Chapter 6 (Szasz et al., 2016).

CHAPTER 9

FUTURE DIRECTIONS

Chapter purpose

This chapter documents the future work that should to be carried out over the short-(two years) and long-(five to ten years) term, based on the findings of the studies contained within this thesis. Short-term goals include 1) the incorporation of qualitative data into performance assessments, and 2) training faculty (internal) assessors to provide appropriate formative feedback to trainees. Long-term goals include 3) the implementation and evaluation of the training and assessment model described in Chapter 4, 4) gathering further evidence from other specialties to determine whether internal assessors can be utilized for summative assessments, 5) creating performance standards for other essential procedures in General Surgery, and 6) implementing and evaluating this collection of procedures into a summative assessment in General Surgery.

9 Future Directions

9.1 Incorporation of qualitative data into performance assessments

When assessing technical and non-technical performance, there can be some disparity of the currently utilized quantitative assessment instruments to capture the less obvious and more difficult to measure competencies/domains (Ginsburg, Gold, Cavalcanti, Kurabi, & McDonald-Blumer, 2011; Hawkins et al., 2015; van der Vleuten & Schuwirth, 2005). As a result, there have been proponents of incorporating more qualitative or narrative components alongside these quantitative instruments during both formative and summative assessments (Ginsburg et al., 2013; Ginsburg et al., 2011; Hawkins et al., 2015; van der Vleuten & Schuwirth, 2005). During formative assessments these new components can better pinpoint avenues for improvement, while at the time of summative assessments, they may be able to better discriminate between varying levels of performance (Frohna & Stern, 2005; Ginsburg et al., 2013; Hawkins et al., 2015).

Future work should focus on incorporating such qualitative or narrative components into the already utilized quantitative assessments instruments. Although this seems reasonable for formative assessments, the need to ensure appropriate psychometric properties for use during summative assessments may complicate the use of such qualitative data (Dannefer, 2013; Hawkins et al., 2015; Holmboe et al., 2010). However, modalities have been developed to increase the validity and reliability of narrative comments and rater training should focus on these (Driessen, van der Vleuten, Schuwirth, van Tartwijk, & Vermunt, 2005; van der Vleuten & Schuwirth, 2005). Finally, although some work has been completed in this regard in Internal Medicine, comparisons between the quantitative and qualitative data for the same performance assessment should be compared to see if they are both able to capture the underlying trainee performance (Ginsburg et al., 2013).

9.2 Faculty development in formative feedback

Recent literature has demonstrated that a major barrier to incorporating competency-based assessments into residency training is a lack of faculty development (C. Carraccio et al., 2016; Dath & Iobst, 2010; Holmboe et al., 2011). Training faculty (internal) assessors to provide appropriate feedback to trainees is the next step to our work presented in Chapter 5, and to incorporating competency-based formative assessments into CBME. Although we discussed the facets of, and guidelines for, implementing formative assessments in sections 1.5.4 and 7.4, respectively. The focus here is on developing faculty skills in providing formative feedback specifically.

A formative feedback education curriculum should be developed focusing on faculty (staff surgeons) involved in resident training, at all of the General Surgery residency programs in the Canada. This curriculum can be disseminated amongst all of the other surgical specialties associated with the RCPSC and abroad as well, as these programs are dealing with similar barriers ("ACGME Program Requirements for Graduate Medical Education in General Surgery ", 2015; J. R. Frank, Snell, L.S., Sherbino, J, 2014; "Good Medical Practice," 2014; "A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide," 2014). This curriculum should emphasize the components that will create a meaningful and sustained formative feedback program with faculty buy in (Ramani & Krackov, 2012). It should focus on educating faculty regarding: 1) Where these encounters should occur (in the workplace, be it the OR, clinic, or patient ward) and that performance must be directly observed, not inferred. 2) The purpose of formative assessment/feedback (in contrast to summative assessments) in the learning process of trainees, as well as developing an understanding of the various competencies/domains that can be assessed during different workplace encounters (D. A. Cook et al., 2014; M. Feldman et al., 2012; "General Surgery," 2015; Ramani & Krackov, 2012). 3) The appropriate timing of feedback, in terms of it occurring soon after the completion of an observed performance, to avoid lag time bias and to provide opportunities for trainee improvement at subsequent formative assessments, and the need for regular ongoing feedback (Gonzalo et al., 2014; Ramani & Krackov, 2012). 4) The components of effective feedback including, beginning with the learner discussing their own performance, reinforcing positive performance elements, correcting negative performance elements, and creating an explicit plan for improvement moving towards the next assessment (Gonzalo et al., 2014; Ramani & Krackov, 2012). 5) How to actually use the

instruments, turn observable performances into assessments, and subsequently structure trainee feedback (Ramani & Krackov, 2012; Weitz et al., 2014). After the incorporation of this curriculum, the abilities of the faculty to provide appropriate feedback should also be assessed and calibrated, providing the faculty with advice on how these abilities can be improved, as literature has shown that faculty are poor assessors of their own assessment abilities (Eva & Regehr, 2008; Holmboe et al., 2011).

9.3 Implementation and evaluation of the training and assessment model

The next step for the education and assessment model discussed in Chapter 4 is to implement it into General Surgery training in Canada and perhaps the United States, replacing the outdated and non-specific guidelines currently in use ("Curriculum outline for General Surgery Residency," 2015-2016; "The Future of General Surgery: Evolving to meet a changing practice," 2014; "General Surgery: Content Outline for the ABS In-Training Examination," 2013; "Objectives of Training in the Specialty of General Surgery," 2010; Szasz et al., 2016).

Subsequently, once the program (model) has been implemented, the next step is to evaluate it. Program evaluation although a complex topic in its own right, can take on two basic forms, constructive or conclusive (Chen, 2015a). Constructive evaluation is used when a program is relatively new and the goal of the evaluation is to improve the program (Chen, 2015a). Conclusive evaluation is used to determine whether a program has merit and whether it realizes the goals that it was set out to accomplish (whether it is a success or failure) (Chen, 2015a). Furthermore, the evaluation can focus on the processes of the program or its outcomes (Chen, 2015a).

For our training and assessment model, initially, constructive process evaluation should be completed, with the goal of focusing on the processes or structural changes that can be implemented that will help the program grow and improve. This should eventually (five years or so) be followed by a conclusive outcome evaluation, to determine whether this new program has been a success or failure in improving trainee competence and confidence after the completion of General Surgery residency, something the current training guidelines are failing to do (Coleman et al., 2013; "Curriculum outline for General Surgery Residency," 2015-2016; Fonseca et al., 2014; "The Future of General Surgery: Evolving to meet a changing practice," 2014; "General

Surgery: Content Outline for the ABS In-Training Examination," 2013; "Objectives of Training in the Specialty of General Surgery," 2010).

9.4 Collection of evidence to support internal assessors for summative assessments

The study described in Chapter 5 of this thesis, provides evidence to support the use of internal assessors for formative assessments in terms of their accuracy. Adding to this, the ability of these internal assessors based on their underlying characteristics and experiences coupled with appropriate training surrounding feedback, further supports their utility in a formative manner (Gonzalo et al., 2014; Ramani & Krackov, 2012).

Despite this information, the use of internal assessors during summative assessments is inconclusive. Given that such assessments can have severe consequences, one study alone (even if supportive) is insufficient to answer such a complex question (Hawkins et al., 2015; Holmboe et al., 2010). Therefore, further studies must be carried out that will eventually support or refute the use of internal assessors and provide generalizable data. These studies should occur in a variety of surgical and medical specialties across multiple institutions, evaluating an assortment of competencies/domains (J. R. Frank & Danoff, 2007; J. R. Frank, Snell. L., Sherbino, J, 2015).

In General Surgery specifically, procedures in addition to the laparoscopic cholecystectomy should be assessed by a heterogeneous group of internal assessors, in terms of their age, gender, clinical experience, type of practice (academic versus community) and associated University appointments. As well, information about how familiar the internal raters are with the trainees they assess and their previous training/exposure to performance assessments should also be sought, to determine the impact these variables may have on the resultant assessment scores. Finally, although difficult to determine in the OR, and not possible in our study, due to the OR structure at the University of Toronto, multiple internal assessors should evaluate the same trainee during the same case, to provide ideas of assessor reliability (other than those compared to an external rater). This reliability if not feasible in the OR, may be established in a simulated setting, where the constraints of having multiple raters in the same OR for a particular procedure are absent. The information gained in the simulated setting, although not ideal, may lend some information to internal rater reliability in the OR, as has been completed with other studies (Dedy, Szasz, et al., 2015)

9.5 Creation of performance standards for other essential procedures in General Surgery

In our study described in Chapter 6, we demonstrated that it was possible to set credible and reliable standards for technical and non-technical performance in General Surgery, for the laparoscopic cholecystectomy. The information garnered in that study can be used to create performance standards from both a technical and non-technical standpoint for other essential procedures in General Surgery.

Elucidating which procedures are essential to the practice of General Surgery or to different stages of General Surgery training, can in part be based on the work we completed in Chapter 4, delineating procedures for milestone assessments, as well as the work completed by Ten Cate *et al*., Carraccio *et al*., and Aylward *et al.*, documenting the selection of essential (entrustable) procedures in CBME (Aylward, Nixon, & Gladding, 2014; C. Carraccio et al., 2016; ten Cate & Scheele, 2007). These elucidated procedures can then be compared to the literature evaluating common procedural exposures in General Surgery trainees and practicing surgeons to ensure overlap (R. H. Bell, Jr. et al., 2009; M. F. Brennan & Debas, 2004; R. S. Chung, 2005; "The Future of General Surgery: Evolving to meet a changing practice," 2014)

Once the specific procedures have been determined, standards can be methodically set that determine when trainees have achieved an appropriate level of performance, while also ensuring their credibility and reliability (Cizek & Bunch, 2007f; J. J. Norcini, 2003; J. J. Norcini, Shea, JA, 1997). As standards are context and procedure dependent, created for a particular purpose, their incorporation into CBME needs to be specific (Cizek & Bunch, 2007f; Holmboe et al., 2010; J. J. Norcini, 2003; J. J. Norcini, Shea, JA, 1997). They can be implemented to determine resident transition during training (i.e. from a junior to senior resident) as was the case with the four procedures identified as milestones in our assessment model. Alternatively, standards (now for a different set of essential procedures) can be set as an assessment at the completion of training, in addition to the written and oral components of the RCPSC licensing examination ("Specialty Training Requirements in General Surgery," 2015).

The lessons learned in General Surgery by this collection of essential procedures; the setting of standards, and their eventual implementation into CBME and assessment can then go on to serve as a model for other surgical and medical specialties.

9.6 Implementation and evaluation of these procedures as a summative assessment in General Surgery

Building on section 9.5, once standards have been set for various essential procedures and their particular purpose has been determined (i.e. resident transition during training), they should be implemented into CBME and assessment, and evaluated.

It is our view that this collection of procedures (the four identified in our study in Chapter 4, plus perhaps others determined in the manner described above) be implemented as a summative assessment that together needs to be attained from both a technical and non-technical standpoint for trainee progression from a junior to senior trainee. This summative assessment would contain these procedures, observed by multiple raters on multiple occasions to ensure the observed overall performance is an indication of trainee 'true' performance (Crossley et al., 2007; J. R. Wilkinson et al., 2008). Corroborating this to the traditional training framework, this assessment could occur at some point in the PGY3 year of training (but based on the CBME framework, really whenever the trainee is deemed ready). If trainees are successful at this summative assessment, they can proceed into senior resident training. If trainees are not successful, a remediation program should be initiated focusing on the areas needed for improvement (Bhatti et al., 2015; Holmboe et al., 2010; Wu et al., 2010). However, as the CBME framework should have multiple opportunities for formative assessment and feedback during training, residents should not be endorsed to undertake this summative assessment until truly ready (Hawkins et al., 2015; Holmboe et al., 2010). Using trainee case numbers as a predictor of readiness to undertake such assessments as outlined in Chapter 6 can further support this. As a result, those requiring remediation due to a poor performance during a summative assessment should be minimal (true negative trainees).

Once implemented, this collection of procedures should undergo evaluation similar to that described in section 9.3. In this instance, the focus would be solely on the outcomes of these assessments (whether trainees are competent or non-competent). Initially, the evaluation would be constructive, determining ways to improve facets around summative assessment implementation (Chen, 2015a). Subsequently, the evaluation would transition to be conclusive, determining whether this assessment is fulfilling its goals of creating credible and reliable methods to differentiate between appropriate levels of performance (Chen, 2015a). This

conclusive evaluation can also be aided by ongoing research collecting evidence to support the use of these assessment scores in keeping with Messick's conceptual framework of validity and the use of performance standards in keeping with Norcini's framework of credibility (Chen, 2015a; S. M. Downing, Tekian, A, Yudkowsky, R, 2006; S. Messick, 1989; Messick, 1995; J. J. Norcini, Shea, JA, 1997). If this collective summative assessment to determine progression during training is ultimately deemed successful, resources should be allocated to ensure its sustainability, and future research and development should ensure continual improvement.

References

Aagaard, E., Kane, G. C., Conforti, L., Hood, S., Caverzagie, K. J., Smith, C., . . . Iobst, W. F. (2013). Early feedback on the use of the internal medicine reporting milestones in assessment of resident performance. *J Grad Med Educ, 5*(3), 433-438. doi: 10.4300/JGME-D-13-00001.1

About the USMLE (2016). *United States Medical Licensing Examination.* from http://www.usmle.org/about/

ABS Booklet of Information Surgery. (2015) (pp. 1-42). Philadelphia, PA: American Board of Surgery.

Accreditation Council for Graduate Medical Education (ACGME) - Next Accreditation System (NAS) - Milestones. (2015). from http://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx

Accreditation Council for Graduate Medical Education: About. (2016). from https://http://www.acgme.org/acgmeweb/tabid/116/About.aspx

ACGME Program Requirements for Graduate Medical Education in General Surgery (2015) (pp. 1-32). Chicago, IL: Accreditation Council for Graduate Medical Education.

ACS - years of postgraduate training (2016). from https://http://www.facs.org/education/resources/medical-students/faq/training

Admission Requirements of Canadian Faculties of Medicine. (2015) (pp. 1-64). Ottawa, ON: The Association of Faculties of Medicine of Canada.

Adrales, G. L., Chu, U. B., Witzke, D. B., Donnelly, M. B., Hoskins, D., Mastrangelo, M. J., Jr., . . . Park, A. E. (2003). Evaluating minimally invasive surgery training using low-cost mechanical simulations. *Surg Endosc, 17*(4), 580-585. doi: 10.1007/s00464-002-8841-7

Adrales, G. L., Donnelly, M. B., Chu, U. B., Witzke, D. B., Hoskins, J. D., Mastrangelo, M. J., Jr., . . . Park, A. E. (2004). Determinants of competency judgments by experienced laparoscopic surgeons. *Surg Endosc, 18*(2), 323-327. doi: 10.1007/s00464-002-8958-8

Adrales, G. L., Park, A. E., Chu, U. B., Witzke, D. B., Donnelly, M. B., Hoskins, J. D., . . . Gandsas, A. (2003). A valid method of laparoscopic simulation training and competence assessment. *J Surg Res, 114*(2), 156-162.

Aggarwal, R., Darzi, A. (2010). Measurement of Surgical Performance for Delivery of a Comptency-Based Training Curriculum. In T. Athanasiou, Debas, H., Darzi, A (Ed.), *Key Topics in Surgical Research and Methodology* (pp. 115-127). Berlin, Germany: Springer Verlag.

Ahlberg, G., Enochsson, L., Gallagher, A. G., Hedman, L., Hogman, C., McClusky, D. A., 3rd, . . . Arvidsson, D. (2007). Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg, 193*(6), 797-804. doi: 10.1016/j.amjsurg.2006.06.050

Ahlering, T. E., Skarecky, D., Lee, D., & Clayman, R. V. (2003). Successful transfer of open surgical skills to a laparoscopic environment using a robotic interface: initial experience with laparoscopic radical prostatectomy. *J Urol, 170*(5), 1738-1741. doi: 10.1097/01.ju.0000092881.24608.5e

Ahmed, A., Ishman, S. L., Laeeq, K., & Bhatti, N. I. (2013). Assessment of improvement of trainee surgical skills in the operating room for tonsillectomy. *Laryngoscope, 123*(7), 1639-1644. doi: 10.1002/lary.24023

Ahmed, K., Miskovic, D., Darzi, A., Athanasiou, T., & Hanna, G. B. (2011). Observational tools for assessment of procedural skills: a systematic review. *Am J Surg, 202*(4), 469-480 e466. doi: 10.1016/j.amjsurg.2010.10.020

Ahmed, M., Sevdalis, N., Vincent, C., & Arora, S. (2013). Actual vs perceived performance debriefing in surgery: practice far from perfect. *Am J Surg, 205*(4), 434-440. doi: 10.1016/j.amjsurg.2013.01.007

Ahmed, N., Devitt, K. S., Keshet, I., Spicer, J., Imrie, K., Feldman, L., . . . Rutka, J. (2014). A systematic review of the effects of resident duty hour restrictions in surgery: impact on resident wellness, training, and patient outcomes. *Ann Surg, 259*(6), 1041-1053. doi: 10.1097/SLA.0000000000000595

Akhtar, K., Sugand, K., Wijendra, A., Standfield, N. J., Cobb, J. P., & Gupte, C. M. (2015). Training safer surgeons: How do patients view the role of simulation in orthopaedic training? *Patient Saf Surg, 9*, 11. doi: 10.1186/s13037-015-0058-5

Al-Qurayshi, Z., Robins, R., Hauch, A., Randolph, G. W., & Kandil, E. (2016). Association of Surgeon Volume With Outcomes and Cost Savings Following Thyroidectomy: A National Forecast. *JAMA Otolaryngol Head Neck Surg, 142*(1), 32-39. doi: 10.1001/jamaoto.2015.2503

Alaraj, A., Charbel, F. T., Birk, D., Tobin, M., Luciano, C., Banerjee, P. P., . . . Roitberg, B. (2013). Role of cranial and spinal virtual and augmented reality simulation using immersive touch modules in neurosurgical training. *Neurosurgery, 72 Suppl 1*, 115-123. doi: 10.1227/NEU.0b013e3182753093

Ali, J. M. (2013). Getting lost in translation? Workplace based assessments in surgical training. *Surgeon, 11*(5), 286-289. doi: 10.1016/j.surge.2013.03.001

Allal, L., Lopez, L.M. (2005). Formative Assessment of Learning: A Review of Publications in French *Formative Assessment - Improving Learning in Secondary Classrooms* (pp. 241-264). Paris, FR: Organization for Economic Co-Operation and Development.

Alman, B. A., Ferguson, P., Kraemer, W., Nousiainen, M. T., & Reznick, R. K. (2013). Competency-based education: a new model for teaching orthopaedics. *Instr Course Lect, 62*, 565-569.

The American Board of Surgery: About Us. (2015). from http://www.absurgery.org/default.jsp?abouthome

Anaesthetists' Non-Technical Skills (ANTS) System Handbook v1.0. (2012) *Framework for Observing and Rating Anaesthetists' Non-Technical Skills* (Vol. 1). Aberdeen, United Kingdom: University of Aberdeen.

Anderson, C. I., Gupta, R. N., Larson, J. R., Abubars, O. I., Kwiecien, A. J., Lake, A. D., . . . Basson, M. D. (2013). Impact of objectively assessing surgeons' teaching on effective perioperative instructional behaviors. *JAMA Surg, 148*(10), 915-922. doi: 10.1001/jamasurg.2013.2144

Andrew, S. E., Oswald, A., & Stobart, K. (2014). Bridging the continuum: Analysis of the alignment of undergraduate and postgraduate accreditation standards. *Med Teach, 36*(9), 804-811. doi: 10.3109/0142159X.2014.910298

Angus, S., Moriarty, J., Nardino, R. J., Chmielewski, A., & Rosenblum, M. J. (2015). Internal Medicine Residents' Perspectives on Receiving Feedback in Milestone Format. *J Grad Med Educ, 7*(2), 220-224. doi: 10.4300/JGME-D-14-00446.1

Ansell, J. S., Boughton, R., Cullen, T., Hodges, C., Nation, E., Peters, P., & Scardino, P. (1979). Lack of agreement between subjective ratings of instructors and objective testing of knowledge acquisition in a urological continuing medical education course. *J Urol, 122*(6), 721-723.

Apramian, T., Cristancho, S., Watling, C., Ott, M., & Lingard, L. (2015). Thresholds of Principle and Preference: Exploring Procedural Variation in Postgraduate Surgical Education. *Acad Med, 90*(11 Suppl), S70-76. doi: 10.1097/ACM.0000000000000909

Aquina, C. T., Probst, C. P., Kelly, K. N., Iannuzzi, J. C., Noyes, K., Fleming, F. J., & Monson, J. R. (2015). The pitfalls of inguinal herniorrhaphy: Surgeon volume matters. *Surgery, 158*(3), 736-746. doi: 10.1016/j.surg.2015.03.058

Arora, M., Diwan, A. D., & Harris, I. A. (2013). Burnout in orthopaedic surgeons: a review. *ANZ J Surg, 83*(7-8), 512-515. doi: 10.1111/ans.12292

Asch, D. A., & Parker, R. M. (1988). The Libby Zion case. One step forward or two steps backward? *N Engl J Med, 318*(12), 771-775. doi: 10.1056/NEJM198803243181209

Asimakopoulos, G., Karagounis, A. P., Valencia, O., Rose, D., Niranjan, G., & Chandrasekaran, V. (2006). How safe is it to train residents to perform off-pump coronary artery bypass surgery? *Ann Thorac Surg, 81*(2), 568-572. doi: 10.1016/j.athoracsur.2005.07.054

Awad, S. S., Liscum, K. R., Aoki, N., Awad, S. H., & Berger, D. H. (2002). Does the subjective evaluation of medical student surgical knowledge correlate with written and oral exam performance? *J Surg Res, 104*(1), 36-39. doi: 10.1006/jsre.2002.6401

Aylward, M., Nixon, J., & Gladding, S. (2014). An entrustable professional activity (EPA) for handoffs as a model for EPA assessment development. *Acad Med, 89*(10), 1335-1340. doi: 10.1097/ACM.0000000000000317

Aziz, F. (2015). Vascular surgery trainees still need to learn how to sew: importance of learning surgical techniques in the era of endovascular surgery. *Front Surg, 2*, 16. doi: 10.3389/fsurg.2015.00016

Babineau, T. J., Becker, J., Gibbons, G., Sentovich, S., Hess, D., Robertson, S., & Stone, M. (2004). The "cost" of operative training for surgical residents. *Arch Surg, 139*(4), 366-369; discussion 369-370. doi: 10.1001/archsurg.139.4.366

Baker, J., Misra, S., Manimala, N. J., Kuy, S., & Gantt, G. (2013). The role of politics in shaping surgical training. *Bull Am Coll Surg, 98*(8), 17-25.

Bandiera, G., Maniate, J., Hanson, M. D., Woods, N., & Hodges, B. (2015). Access and Selection: Canadian Perspectives on Who Will Be Good Doctors and How to Identify Them. *Acad Med, 90*(7), 946-952. doi: 10.1097/ACM.0000000000000683

Barden, C. B., Specht, M. C., McCarter, M. D., Daly, J. M., & Fahey, T. J., 3rd. (2002). Effects of limited work hours on surgical training. *J Am Coll Surg, 195*(4), 531-538.

Barger, L. K., Ayas, N. T., Cade, B. E., Cronin, J. W., Rosner, B., Speizer, F. E., & Czeisler, C. A. (2006). Impact of extended-duration shifts on medical errors, adverse events, and attentional failures. *PLoS Med, 3*(12), e487. doi: 10.1371/journal.pmed.0030487

Barsuk, J. H., Cohen, E. R., Caprio, T., McGaghie, W. C., Simuni, T., & Wayne, D. B. (2012). Simulation-based education with mastery learning improves residents' lumbar puncture skills. *Neurology, 79*(2), 132-137. doi: 10.1212/WNL.0b013e31825dd39d

Bartlett, K. W., Whicker, S. A., Bookman, J., Narayan, A. P., Staples, B. B., Hering, H., & McGann, K. A. (2015). Milestone-Based Assessments Are Superior to Likert-Type Assessments in Illustrating Trainee Progression. *J Grad Med Educ, 7*(1), 75-80. doi: 10.4300/JGME-D-14-00389.1

Bartman, I., Smee, S.M., Roy, M. (2011). Catching the Hawks and Doves: A Method for Identifying Extreme Examiners on Objective Structured Clinical Examinations. In M. C. o. C. (MCC) (Ed.). Ottawa, Canada: Medical Council of Canada (MCC).

Bath, A. P., & Wilson, T. (2007). Objective assessment of surgical competency--ENT trainees. *Clin Otolaryngol, 32*(6), 475-479. doi: 10.1111/j.1749-4486.2007.01555.x

Beard, J., Rowley, D., Bussey, M., & Pitts, D. (2009). Workplace-based assessment: assessing technical skill throughout the continuum of surgical training. *ANZ J Surg, 79*(3), 148-153. doi: 10.1111/j.1445-2197.2008.04832.x

Beard, J. D. (2008). Assessment of surgical skills of trainees in the UK. *Ann R Coll Surg Engl, 90*(4), 282-285. doi: 10.1308/003588408X286017

Beard, J. D., Choksy, S., & Khan, S. (2007). Assessment of operative competence during carotid endarterectomy. *Br J Surg, 94*(6), 726-730. doi: 10.1002/bjs.5689

Beard, J. D., Education, Training Committee of the Vascular Society of Great, B., & Ireland. (2005). Setting standards for the assessment of operative competence. *Eur J Vasc Endovasc Surg, 30*(2), 215-218. doi: 10.1016/j.ejvs.2005.01.032

Bell, R. H., Jr. (2009). Why Johnny cannot operate. *Surgery, 146*(4), 533-542. doi: 10.1016/j.surg.2009.06.044

Bell, R. H., Jr., Biester, T. W., Tabuenca, A., Rhodes, R. S., Cofer, J. B., Britt, L. D., & Lewis, F. R., Jr. (2009). Operative experience of residents in US general surgery programs: a gap between expectation and experience. *Ann Surg, 249*(5), 719-724. doi: 10.1097/SLA.0b013e3181a38e59

Bell, R. M., Fann, S. A., Morrison, J. E., & Lisk, J. R. (2011). Determining personal talents and behavioral styles of applicants to surgical training: a new look at an old problem, part I. *J Surg Educ, 68*(6), 534-541. doi: 10.1016/j.jsurg.2011.05.016

Benefits of CBD for medical educators in each specialty. (2014). *Competence by Design (CBD)*. from http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme

Berkey, T. (1994). Benchmarking in health care: turning challenges into success. *Jt Comm J Qual Improv, 20*(5), 277-284.

Berry, W. (2006). Surgical malpractice: myths and realities. *J Med Pract Manage, 22*(1), 50-51.

Berwick, D. M. (1989). Continuous improvement as an ideal in health care. *N Engl J Med, 320*(1), 53-56. doi: 10.1056/NEJM198901053200110

Bhandari, M., Montori, V., Devereaux, P. J., Dosanjh, S., Sprague, S., & Guyatt, G. H. (2003). Challenges to the practice of evidence-based medicine during residents' surgical training: a qualitative study using grounded theory. *Acad Med, 78*(11), 1183-1190.

Bhatti, N. I., Ahmed, A., Stewart, M. G., Miller, R. H., & Choi, S. S. (2015). Remediation of problematic residents-A national survey. *Laryngoscope*. doi: 10.1002/lary.25599

Bhatti, N. I., & Cummings, C. W. (2007). Competency in surgical residency training: defining and raising the bar. *Acad Med, 82*(6), 569-573. doi: 10.1097/ACM.0b013e3180555bfb

Bindal, T., Wall, D., & Goodyear, H. M. (2011). Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise? *Med Teach, 33*(11), 919-927. doi: 10.3109/0142159X.2011.558140

Birkmeyer, J. D., Finks, J. F., O'Reilly, A., Oerline, M., Carlin, A. M., Nunn, A. R., . . . Michigan Bariatric Surgery, C. (2013). Surgical skill and complication rates after bariatric surgery. *N Engl J Med, 369*(15), 1434-1442. doi: 10.1056/NEJMsa1300625

Birkmeyer, J. D., Siewers, A. E., Finlayson, E. V., Stukel, T. A., Lucas, F. L., Batista, I., . . . Wennberg, D. E. (2002). Hospital volume and surgical mortality in the United States. *N Engl J Med, 346*(15), 1128-1137. doi: 10.1056/NEJMsa012337

Bland, J. M., & Altman, D. G. (1997). Cronbach's alpha. *BMJ, 314*(7080), 572.

Blum, A. B., Raiszadeh, F., Shea, S., Mermin, D., Lurie, P., Landrigan, C. P., & Czeisler, C. A. (2010). US public opinion regarding proposed limits on resident physician work hours. *BMC Med, 8*, 33. doi: 10.1186/1741-7015-8-33

Blum, C. A., & Adams, D. B. (2011). Who did the first laparoscopic cholecystectomy? *J Minim Access Surg, 7*(3), 165-168. doi: 10.4103/0972-9941.83506

Blumberg, P., & Michael, J.A. (1992). Development of self‐directed learning behaviors in a partially teacher‐directed problem‐based learning curriculum. *Teach Learn Med, 3*(1), 3-8. doi: 10.1080/10401339209539526

Boet, S., Bould, M. D., Fung, L., Qosa, H., Perrier, L., Tavares, W., . . . Tricco, A. C. (2014). Transfer of learning and patient outcome in simulated crisis resource management: a systematic review. *Can J Anaesth, 61*(6), 571-582. doi: 10.1007/s12630-014-0143-8

Bonrath, E. M., Dedy, N. J., Gordon, L. E., & Grantcharov, T. P. (2015). Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Ann Surg, 262*(2), 205-212. doi: 10.1097/SLA.0000000000001214

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality assurance methods for performance-based assessments. *Adv Health Sci Educ Theory Pract, 8*(1), 27-47.

Brennan, M. F., & Debas, H. T. (2004). Surgical education in the United States: portents for change. *Ann Surg, 240*(4), 565-572.

Brennan, R. L., Lockwood, R.E. (1980). A Comparison of the Nedelsky and Angoff Cutting Score Procedures Using Generalizability Theory. *Appl Psych Meas, 4*(2), 219-240.

Bridges, M., & Diamond, D. L. (1999). The financial impact of teaching surgical residents in the operating room. *Am J Surg, 177*(1), 28-32.

Brindley, P. G., Jones, D. B., Grantcharov, T., & de Gara, C. (2012). Canadian Association of University Surgeons' Annual Symposium. Surgical simulation: the solution to safe training or a promise unfulfilled? *Can J Surg, 55*(4), S200-206. doi: 10.1503/cjs.027910

Brinkman, W. M., Buzink, S. N., Alevizos, L., de Hingh, I. H., & Jakimowicz, J. J. (2012). Criterion-based laparoscopic training reduces total training time. *Surg Endosc, 26*(4), 1095-1101. doi: 10.1007/s00464-011-2005-6

Brunner, W. C., Korndorffer, J. R., Jr., Sierra, R., Dunne, J. B., Yau, C. L., Corsetti, R. L., . . . Scott, D. J. (2005). Determining standards for laparoscopic proficiency using virtual reality. *Am Surg, 71*(1), 29-35.

Brydges, R., Classen, R., Larmer, J., Xeroulis, G., & Dubrowski, A. (2006). Computer-assisted assessment of one-handed knot tying skills performed within various contexts: a construct validity study. *Am J Surg, 192*(1), 109-113. doi: 10.1016/j.amjsurg.2005.11.014

Bufalari, A., Ferri, M., Cao, P., Cirocchi, R., Bisacci, R., & Moggi, L. (1996). Surgical care in octogenarians. *Br J Surg, 83*(12), 1783-1787.

Camp, R. C. (1992). Learning from the best leads to superior performance. *J Bus Strategy, 13*(3), 3-6.

Canadian Resident Matching Service (CaRMS) - examinations and assessments (2015). from http://www.carms.ca/en/match-process/your-application/documents/examinations-assessments/

CanMEDS. (2013). *Royal College of Physicians and Surgeons of Canada.* Retrieved August 15, 2013, from http://www.royalcollege.ca/portal/page/portal/rc/canmeds

CanMEDS 2005 Framework. (2005) (1st ed., pp. 1-11). Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

CanMEDS history. (2015). from http://www.royalcollege.ca/portal/page/portal/rc/canmeds/about/history

Carlsen, C. G., Lindorff-Larsen, K., Funch-Jensen, P., Lund, L., Morcke, A. M., Ipsen, M., & Charles, P. (2014). Is current surgical training efficient? A national survey. *J Surg Educ, 71*(3), 367-374. doi: 10.1016/j.jsurg.2013.10.002

Carlson, M. L., Archibald, D. J., Sorom, A. J., & Moore, E. J. (2010). Under the microscope: assessing surgical aptitude of otolaryngology residency applicants. *Laryngoscope, 120*(6), 1109-1113.

Carraccio, C., Englander, R., Gilhooly, J., Mink, R., Hofkosh, D., Barone, M. A., & Holmboe, E. S. (2016). Building a Framework of Entrustable Professional Activities, Supported by Competencies and Milestones, to Bridge the Educational Continuum. *Acad Med*. doi: 10.1097/ACM.0000000000001141

Carraccio, C., Englander, R., Van Melle, E., Ten Cate, O., Lockyer, J., Chan, M. K., . . . International Competency-Based Medical Education, C. (2015). Advancing Competency-Based Medical Education: A Charter for Clinician-Educators. *Acad Med*. doi: 10.1097/ACM.0000000000001048

Carraccio, C., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms: from Flexner to competencies. *Acad Med, 77*(5), 361-367.

Carraccio, C. L., & Englander, R. (2013). From Flexner to competencies: reflections on a decade and the journey ahead. *Acad Med, 88*(8), 1067-1073. doi: 10.1097/ACM.0b013e318299396f

Carthey, J., de Leval, M.R., Wright, D.J., Farewell, V.T., Reason, J.T. (2003). Behavioural markers of surgical excellence. *Safety Sci, 41*, 409-425.

Champion, H. R., & Gallagher, A. G. (2003). Surgical simulation - a 'good idea whose time has come'. *Br J Surg, 90*(7), 767-768. doi: 10.1002/bjs.4187

Chen, H. T. (2015a). Fundamentals of Program Evaluation. In H. T. Chen (Ed.), *Practical Program Evaluation: Theory-Driven Evaluation and the Integrated Evaluation Perspective* (2nd ed., pp. 3-33). Thousand Oaks, CA: SAGE Publications.

Chen, H. T. (2015b). Logic Models and the Action Model/Change Model Schema (Program Theory). In H. T. Chen (Ed.), *Practical Program Evaluation: Theory-Driven Evaluation and the Integrated Evaluation Perspective* (2nd ed., pp. 58-93). Thousand Oaks, CA: SAGE Publications.

Chipman, J. G., & Schmitz, C. C. (2009). Using objective structured assessment of technical skills to evaluate a basic skills simulation curriculum for first-year surgical residents. *J Am Coll Surg, 209*(3), 364-370 e362. doi: 10.1016/j.jamcollsurg.2009.05.005

Chiu, C. C., Wei, P. L., Wang, W., Chen, R. J., Chen, T. C., Lee, W. J., & Huang, M. T. (2006). Role of appendectomy in laparoscopic training. *J Laparoendosc Adv Surg Tech A, 16*(2), 113-118. doi: 10.1089/lap.2006.16.113

Chou, B., Bowen, C. W., & Handa, V. L. (2008). Evaluating the competency of gynecology residents in the operating room: validation of a new assessment tool. *Am J Obstet Gynecol, 199*(5), 571 e571-575. doi: 10.1016/j.ajog.2008.06.082

Christein, J. D., Cook, J. K., Enger, T. M., & Farley, D. R. (2006). Preliminary general surgery residents: indentured servitude or golden opportunity? *Curr Surg, 63*(1), 85-89. doi: 10.1016/j.cursur.2005.10.001

Chung, R., Pham, Q., Wojtasik, L., Chari, V., & Chen, P. (2003). The laparoscopic experience of surgical graduates in the United States. *Surg Endosc, 17*(11), 1792-1795. doi: 10.1007/s00464-002-8922-7

Chung, R. S. (2005). How much time do surgical residents need to learn operative surgery? *Am J Surg, 190*(3), 351-353. doi: 10.1016/j.amjsurg.2005.06.035

Churchill, E. D. (1939). Graduate training at the Massachusetts General Hospital: a report to the trustees from the general executive committee. Boston.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93-106.

Cizek, G. J., & Bunch, M. B. (2007a). The Angoff Method and Angoff Variations. In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 81-96). Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007b). Common Elements in Setting Performance Standards. In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 34-68). Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007c). Contemporary Standard Setting: An Enduring Need. In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 4-13). Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007d). The Contrasting Groups and Borderline Group Methods. In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 105-117). Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007e). The Hofstee and Beuk Methods. In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 206-219). Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., & Bunch, M. B. (2007f). What is Standard Setting? In G. J. Cizek, Bunch, M.B (Ed.), Standard Setting (pp. 13-34). Thousand Oaks, CA: SAGE Publications.

Cogbill, T., et al. (2014). The General Surgery Milestone Project *The Accreditation Council for Gradaute Medical Education and The American Board of Surgery*: The Accreditation Council for Gradaute Medical Education and The American Board of Surgery.

Cohen, E. R., Barsuk, J. H., McGaghie, W. C., & Wayne, D. B. (2013). Raising the bar: reassessing standards for procedural competence. *Teach Learn Med, 25*(1), 6-9. doi: 10.1080/10401334.2012.741540

Coleman, J. J., Esposito, T. J., Rozycki, G. S., & Feliciano, D. V. (2013). Early subspecialization and perceived competence in surgical training: are residents ready? *J Am Coll Surg, 216*(4), 764-771; discussion 771-763. doi: 10.1016/j.jamcollsurg.2012.12.045

Collar, R. M., Shuman, A. G., Feiner, S., McGonegal, A. K., Heidel, N., Duck, M., . . . Bradford, C. R. (2012). Lean management in academic surgery. *J Am Coll Surg, 214*(6), 928-936. doi: 10.1016/j.jamcollsurg.2012.03.002

A collective vision for postgraduate medical education in Canada. (2012) *The Future of Medical Education in Canada* (pp. 1-48). Ottawa, ON: The Future of Medical Education in Canada.

Common Program Requirements (2016) (pp. 1-24). Chicago, IL: Accreditation Council for Gradaute Medical Education.

Compeau, C., Tyrwhitt, J., Shargall, Y., & Rotstein, L. (2009). A retrospective review of general surgery training outcomes at the University of Toronto. *Can J Surg, 52*(5), E131-136.

Competence by Design (CBD): Moving towards competency-based medical education. (2014). from http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme

Competency-based medical education. (2013). *Royal College of Physicians and Surgeons of Canada.* Retrieved September 1, 2013, from http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme

Conway, N. E., Seymour, N.E., Bush, R.W., Romanelli, J.R., . (2011). Surgical resident learnin curve for a simulated single port laparoscopic surgical task. . *Surg Endosc, 25S*, 235–S240.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med, 119*(2), 166 e167-116. doi: 10.1016/j.amjmed.2005.10.036

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med, 24*(1), 74-79. doi: 10.1007/s11606-008-0842-3

Cook, D. A., & Lineberry, M. (2016). Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med*. doi: 10.1097/ACM.0000000000001114

Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ Theory Pract, 19*(2), 233-250. doi: 10.1007/s10459-013-9458-4

Cook, M. R., Watters, J. M., Barton, J. S., Kamin, C., Brown, S. N., Deveney, K. E., & Kiraly, L. N. (2015). A flexible postoperative debriefing process can effectively provide formative resident feedback. *J Am Coll Surg, 220*(5), 959-967. doi: 10.1016/j.jamcollsurg.2014.12.048

Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology, 78*(1), 98-104.

Crank-Patton, A., Fisher, J. B., & Toedter, L. J. (2001). The role of the journal club in surgical residency programs: a survey of APDS program directors. *Curr Surg, 58*(1), 101-104.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Med Educ, 46*(1), 28-37. doi: 10.1111/j.1365-2923.2011.04166.x

Crossley, J., Marriott, J., Purdie, H., & Beard, J. D. (2011). Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre. *Br J Surg, 98*(7), 1010-1020. doi: 10.1002/bjs.7478

Crossley, J., Russell, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., & Norcini, J. (2007). 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ, 41*(10), 926-934. doi: 10.1111/j.1365-2923.2007.02843.x

Cuhls, K. (2005). Delphi methods : UNIDO Foresight Seminars. Retrieved August 5, 2014, 2014, from http://www.unido.org/fileadmin/import/16959_DelphiMethod.pdf

Curriculum outline for General Surgery Residency. (2013-2014) *Surgical Council on Resident Education*. Philadelphia, PA: Surgical Council on Resident Education.

Curriculum outline for General Surgery Residency. (2015-2016) *Surgical Council on Resident Education*. Philadelphia, PA: Surgical Council on Resident Education.

Cushchieri, A. (2003). Lest we forget the surgeon. *Semin Laparosc Surg, 10*(3), 141-148.

Dannefer, E. F. (2013). Beyond assessment of learning toward assessment for learning: educating tomorrow's physicians. *Med Teach, 35*(7), 560-563. doi: 10.3109/0142159X.2013.787141

Darzi, A. (2008). Quality and the NHS next stage review. *Lancet, 371*(9624), 1563-1564. doi: 10.1016/S0140-6736(08)60672-8

Darzi, A., Datta, V., & Mackay, S. (2001). The challenge of objective assessment of surgical skill. *Am J Surg, 181*(6), 484-486.

Dath, D., & Iobst, W. (2010). The importance of faculty development in the transition to competency-based medical education. *Med Teach, 32*(8), 683-686. doi: 10.3109/0142159X.2010.500710

Dauphinee, W. D., Blackmore, D. E., Smee, S., Rothman, A. I., & Reznick, R. (1997). Using the Judgments of Physician Examiners in Setting the Standards for a National Multi-center High Stakes OSCE. *Adv Health Sci Educ Theory Pract, 2*(3), 201-211. doi: 10.1023/A:1009768127620

de Montbrun, S., Roberts, P. L., Satterthwaite, L., & MacRae, H. (2016). Implementing and Evaluating a National Certification Technical Skills Examination: The Colorectal Objective Structured Assessment of Technical Skill. *Ann Surg*. doi: 10.1097/SLA.0000000000001620

de Montbrun, S., Satterthwaite, L., & Grantcharov, T. P. (2016). Setting pass scores for assessment of technical performance by surgical trainees. *Br J Surg, 103*(3), 300-306. doi: 10.1002/bjs.10047

de Montbrun, S. L., Roberts, P. L., Lowry, A. C., Ault, G. T., Burnstein, M. J., Cataldo, P. A., . . . MacRae, H. (2013). A novel approach to assessing technical competence of colorectal surgery residents: the development and evaluation of the Colorectal Objective Structured Assessment of Technical Skill (COSATS). *Ann Surg, 258*(6), 1001-1006. doi: 10.1097/SLA.0b013e31829b32b8

de Montbrun, S. L., Satterthwaite, L.M., Grantcharov, T.P. (2015). Passing the OSATS exam predicts future technical skill of general surgery residents: Validating the OSATS passing scores. *Manuscipt in preparation*.

de Villiers, M. R., de Villiers, P. J., & Kent, A. P. (2005). The Delphi technique in health sciences education research. *Med Teach, 27*(7), 639-643. doi: 10.1080/13611260500069947

Debas, H. T., Bass, B. L., Brennan, M. F., Flynn, T. C., Folse, J. R., Freischlag, J. A., . . . American Surgical Association Blue Ribbon, C. (2005). American Surgical Association Blue Ribbon Committee Report on Surgical Education: 2004. *Ann Surg, 241*(1), 1-8.

Debes, A. J., Aggarwal, R., Balasundaram, I., & Jacobsen, M. B. (2012). Construction of an evidence-based, graduated training curriculum for D-box, a webcam-based laparoscopic basic skills trainer box. *Am J Surg, 203*(6), 768-775. doi: 10.1016/j.amjsurg.2011.07.022

Dedy, N. J., Bonrath, E. M., Zevin, B., & Grantcharov, T. P. (2013). Teaching nontechnical skills in surgical residency: a systematic review of current approaches and outcomes. *Surgery, 154*(5), 1000-1008. doi: 10.1016/j.surg.2013.04.034

Dedy, N. J., Fecso, A. B., Szasz, P., Bonrath, E. M., & Grantcharov, T. P. (2015). Implementation of an Effective Strategy for Teaching Nontechnical Skills in the Operating Room: A Single-blinded Nonrandomized Trial. *Ann Surg*. doi: 10.1097/SLA.0000000000001297

Dedy, N. J., Szasz, P., Louridas, M., Bonrath, E. M., Husslein, H., & Grantcharov, T. P. (2015). Objective structured assessment of nontechnical skills: Reliability of a global rating scale for the in-training assessment in the operating room. *Surgery, 157*(6), 1002-1013. doi: 10.1016/j.surg.2014.12.023

Del Bigio, M. R. (2007). Please slow down the CanMEDS express. *CMAJ, 176*(6), 812. doi: 10.1503/cmaj.1060218

Directive 2003/88/ES of the European Parliament and of the Council of 4 November 2003 concerning aspects of the organisation of working time. (2003) (pp. 1-11): The European Parliament and the Council of the European Union.

Diwadkar, G. B., van den Bogert, A., Barber, M. D., & Jelovsek, J. E. (2009). Assessing vaginal surgical skills using video motion analysis. *Obstet Gynecol, 114*(2 Pt 1), 244-251. doi: 10.1097/AOG.0b013e3181af25e6

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. *Med Educ, 37*(9), 830-837.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Med Educ, 38*(9), 1006-1012. doi: 10.1111/j.1365-2929.2004.01932.x

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: what about rater error? *Med Educ, 39*(4), 353-355. doi: 10.1111/j.1365-2929.2005.02138.x

Downing, S. M., Tekian, A, Yudkowsky, R. (2006). RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. *Teaching and Learning in Medicine: An International Journal, 18*(1), 50-57.

Drake, F. T., Horvath, K. D., Goldin, A. B., & Gow, K. W. (2013). The general surgery chief resident operative experience: 23 years of national ACGME case logs. *JAMA Surg, 148*(9), 841-847. doi: 10.1001/jamasurg.2013.2919

Dreyfus, S., E., Dreyfus, H,L.,. (1980). A five-stage model of the mental activities involved in directed skill acquisition (pp. 1-22): University of California - Berkeley.

Driessen, E., van der Vleuten, C., Schuwirth, L., van Tartwijk, J., & Vermunt, J. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Med Educ, 39*(2), 214-220. doi: 10.1111/j.1365-2929.2004.02059.x

Drolet, B. C., Sangisetty, S., Tracy, T. F., & Cioffi, W. G. (2013). Surgical residents' perceptions of 2011 Accreditation Council for Graduate Medical Education duty hour regulations. *JAMA Surg, 148*(5), 427-433. doi: 10.1001/jamasurg.2013.169

Dudek, N., & Dojeiji, S. (2014). Twelve tips for completing quality in-training evaluation reports. *Med Teach, 36*(12), 1038-1042. doi: 10.3109/0142159X.2014.932897

Dwyer, T., Wright, S., Kulasegaram, K. M., Theodoropoulos, J., Chahal, J., Wasserstein, D., . . . Ogilvie-Harris, D. (2016). How to set the bar in competency-based medical education: standard setting after an Objective Structured Clinical Examination (OSCE). *BMC Med Educ, 16*(1), 1. doi: 10.1186/s12909-015-0506-z

Eardley, I., Bussey, M., Woodthorpe, A., Munsch, C., & Beard, J. (2013). Workplace-based assessment in surgical training: experiences from the Intercollegiate Surgical Curriculum Programme. *ANZ J Surg, 83*(6), 448-453. doi: 10.1111/ans.12187

Edelman, D. A., Mattos, M. A., & Bouwman, D. L. (2010). FLS skill retention (learning) in first year surgery residents. *J Surg Res, 163*(1), 24-28. doi: 10.1016/j.jss.2010.03.057

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Kit of Factor-Referenced Cognitive Tests. In N. J. Education Testing Service Princeton (Ed.), (pp. p1-314). New Jersey: Office of Naval Research Contract.

Elfenbein, D. M., Sippel, R. S., McDonald, R., Watson, T., Scarborough, J. E., & Migaly, J. (2015). Faculty evaluations of resident medical knowledge: can they be used to predict American Board of Surgery In-Training Examination performance? *Am J Surg, 209*(6), 1095-1101. doi: 10.1016/j.amjsurg.2014.08.042

Elliot, D. L., & Hickam, D. H. (1987). Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. *JAMA, 258*(23), 3405-3408.

Entry Requirements for UK medical schools. (2015) (pp. 1-66). London, UK: Medical Schools Council.

Epstein, R. M. (2007). Assessment in medical education. *N Engl J Med, 356*(4), 387-396. doi: 10.1056/NEJMra054784

Eva, K. W., & Regehr, G. (2008). "I'll never play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof, 28*(1), 14-19. doi: 10.1002/chp.150

Fabricant, P. D., Dy, C. J., Dare, D. M., & Bostrom, M. P. (2013). A narrative review of surgical resident duty hour limits: where do we go from here? *J Grad Med Educ, 5*(1), 19-24. doi: 10.4300/JGME-D-12-00081.1

Falleti, M. G., Maruff, P., Collie, A., Darby, D. G., & McStephen, M. (2003). Qualitative similarities in cognitive impairment associated with 24 h of sustained wakefulness and a blood alcohol concentration of 0.05%. *J Sleep Res, 12*(4), 265-274.

Fargo, M. V., Edwards, J. A., Roth, B. J., & Short, M. W. (2011). Using a simulated surgical skills station to assess laceration management by surgical and nonsurgical residents. *J Grad Med Educ, 3*(3), 326-331. doi: 10.4300/JGME-D-10-00208.1

Farrell, T. M., Kohn, G. P., Owen, S. M., Meyers, M. O., Stewart, R. A., & Meyer, A. A. (2010). Low correlation between subjective and objective measures of knowledge on surgery clerkships. *J Am Coll Surg, 210*(5), 680-683, 683-685. doi: 10.1016/j.jamcollsurg.2009.12.020

Faulkner, H., Regehr, G., Martin, J., & Reznick, R. (1996). Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med, 71*(12), 1363-1365.

Faurie, C., & Khadra, M. (2012). Technical competence in surgeons. *ANZ J Surg, 82*(10), 682-690. doi: 10.1111/j.1445-2197.2012.06239.x

Feldman, L. S., Hagarty, S. E., Ghitulescu, G., Stanbridge, D., & Fried, G. M. (2004). Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *J Am Coll Surg, 198*(1), 105-110. doi: 10.1016/j.jamcollsurg.2003.08.020

Feldman, M., Lazzara, E. H., Vanderbilt, A. A., & DiazGranados, D. (2012). Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof, 32*(4), 279-286. doi: 10.1002/chp.21156

Ferguson, P. C., Kraemer, W., Nousiainen, M., Safir, O., Sonnadara, R., Alman, B., & Reznick, R. (2013). Three-year experience with an innovative, modular competency-based curriculum for orthopaedic training. *J Bone Joint Surg Am, 95*(21), e166. doi: 10.2106/JBJS.M.00314

Figert, P. L., Park, A. E., Witzke, D. B., & Schwartz, R. W. (2001). Transfer of training in acquiring laparoscopic skills. *J Am Coll Surg, 193*(5), 533-537.

Fine, B. A., Golden, B., Hannam, R., Morra, D. (2009). Leading Lean: A Canadian Healthcare Leader's Guide. *Healthc Q, 12*(3), 1-10.

Fink, A., Kosecoff, J., Chassin, M., & Brook, R. H. (1984). Consensus methods: characteristics and guidelines for use. *Am J Public Health, 74*(9), 979-983.

Firth-Cozens, J., & Mowbray, D. (2001). Leadership and the quality of care. *Qual Health Care, 10 Suppl 2*, ii3-7.

Fleming, J., Kapoor, K., Sevdalis, N., & Harries, M. (2012). Validation of an operating room immersive microlaryngoscopy simulator. *Laryngoscope, 122*(5), 1099-1103. doi: 10.1002/lary.23240

Fletcher, K. E., Underwood, W., 3rd, Davis, S. Q., Mangrulkar, R. S., McMahon, L. F., Jr., & Saint, S. (2005). Effects of work hour reduction on residents' lives: a systematic review. *JAMA, 294*(9), 1088-1100. doi: 10.1001/jama.294.9.1088

Flin, R., Martin, R., Goeters, K.M., Hormann, H.J., Amalberti, R., Valot, C., Nijhuis, H. (2003). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills. *Hum Factors Aerospace Saf, 3*(2), 95-117.

Flin, R., O'Connor, P, Crichton, M. (2008). Introduction *Safety at the sharp end: a guide to non-technical skills* (pp. 1-16). Hampshire, UK: Ashgate Publishing Limited

Fonseca, A. L., Reddy, V., Longo, W. E., & Gusberg, R. J. (2014). Graduating general surgery resident operative confidence: perspective from a national survey. *J Surg Res, 190*(2), 419-428. doi: 10.1016/j.jss.2014.05.014

Format of the Comprehensive Objective Examination in General Surgery. (2014). from http://www.royalcollege.ca/rc/faces/oracle/webcenter/portalapp/pages/viewDocument.jspx?document_id=TZTEST3RCPSCED002098&_afrLoop=12292838567792346&_afrWindowMode=0&_afrWindowId=f9se3na5m_1 -!%40%40%3F_afrWindowId%3Df9se3na5m_1%26document_id%3DTZTEST3RCPSCED002098%26_afrLoop%3D12292838567792346%26_afrWindowMode%3D0%26_adf.ctrl-state%3Df9se3na5m_17

Francis, H. W., Masood, H., Laeeq, K., & Bhatti, N. I. (2010). Defining milestones toward competency in mastoidectomy using a skills assessment paradigm. *Laryngoscope, 120*(7), 1417-1421. doi: 10.1002/lary.20953

Frank, J. R., & Danoff, D. (2007). The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach, 29*(7), 642-647. doi: 10.1080/01421590701746983

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Med Teach, 32*(8), 631-637. doi: 10.3109/0142159X.2010.500898

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., . . . Harris, K. A. (2010). Competency-based medical education: theory to practice. *Med Teach, 32*(8), 638-645. doi: 10.3109/0142159X.2010.501190

Frank, J. R., Snell, L.S., Sherbino, J. (2014). Draft CanMEDS 2015 Milestones Guide. In J. R. Frank, Snell, L.S., Sherbino, J. et al (Ed.), *Royal College of Physicians and Surgeons of Canada.* (pp. 1-52). Ottawa, Canada: Royal College of Physicians and Surgeons of Canada.

Frank, J. R., Snell. L., Sherbino, J. (2015). CanMEDS 2015 Physician Competency Framework (pp. 1-17). Ottawa, ON: Royal College of Physicians and Surgeons of Canada

Frank, J. R. J., M. et al. . (2005). Report of the CanMEDS Phase IV Working Groups. Ottawa: The Royal College of Physicians and Surgeons of Canada.

Fraser, S. A., Klassen, D. R., Feldman, L. S., Ghitulescu, G. A., Stanbridge, D., & Fried, G. M. (2003). Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. *Surg Endosc, 17*(6), 964-967. doi: 10.1007/s00464-002-8828-4

Fried, M. P., Kaye, R. J., Gibber, M. J., Jackman, A. H., Paskhover, B. P., Sadoughi, B., . . . Jacobs, J. B. (2012). Criterion-based (proficiency) training to improve surgical performance. *Arch Otolaryngol Head Neck Surg, 138*(11), 1024-1029. doi: 10.1001/2013.jamaoto.377

Frohna, A., & Stern, D. (2005). The nature of qualitative comments in evaluating professionalism. *Med Educ, 39*(8), 763-768. doi: 10.1111/j.1365-2929.2005.02234.x

The Future of General Surgery: Evolving to meet a changing practice. (2014) *Final Report of the Task Force on the Future of General Surgery* (pp. 1- 59). Ottawa, Canada: Royal College of Physicians and Surgeons of Canada.

Gallagher, A. G., Cowie, R., Crothers, I., Jordan-Black, J. A., & Satava, R. M. (2003). PicSOr: an objective test of perceptual skill that predicts laparoscopic technical skill in three initial studies of laparoscopic performance. *Surg Endosc, 17*(9), 1468-1471. doi: 10.1007/s00464-002-8569-4

Gallagher, A. G., Ritter, E. M., Champion, H., Higgins, G., Fried, M. P., Moses, G., . . . Satava, R. M. (2005). Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg, 241*(2), 364-372.

Ganju, A., Kahol, K., Lee, P., Simonian, N., Quinn, S. J., Ferrara, J. J., & Batjer, H. H. (2012). The effect of call on neurosurgery residents' skills: implications for policy regarding resident call periods. *J Neurosurg, 116*(3), 478-482. doi: 10.3171/2011.9.JNS101406

Gatcliffe, T. A., & Coleman, R. L. (2008). Tumor board: more than treatment planning--a 1-year prospective survey. *J Cancer Educ, 23*(4), 235-237. doi: 10.1080/08858190802189014

Gauger, P. G., Hauge, L. S., Andreatta, P. B., Hamstra, S. J., Hillard, M. L., Arble, E. P., . . . Minter, R. M. (2010). Laparoscopic simulation training with proficiency targets improves

practice and performance of novice surgeons. *Am J Surg, 199*(1), 72-80. doi: 10.1016/j.amjsurg.2009.07.034

General Standards Applicable to All Residency Programs: B Standards. (2013) (pp. 1-16). Ottawa, ON: Royal College of Physicians and Surgeons of Canada

General Surgery. (2015) *The Intercollegiate Surgical Curriculum: Educating the surgeons of the future* (pp. 1-336): Intercollegiate Surgical Curriculum Programme.

General Surgery - Canadian Resident Matching Service (CaRMS). (2015). from https://phx.e-carms.ca/phoenix-web/pd/main?mitid=1241

General Surgery Profile. (2015) *Canadian Specialty Profiles*. Ottawa, ON: Canadian Medical Association.

. General Surgery: Content Outline for the ABS In-Training Examination. (2013) (pp. 1-2). Philadelphia, PA: American Board of Surgery.

Ghaderi, I., Manji, F., Park, Y. S., Juul, D., Ott, M., Harris, I., & Farrell, T. M. (2015). Technical skills assessment toolbox: a review using the unitary framework of validity. *Ann Surg, 261*(2), 251-262. doi: 10.1097/SLA.0000000000000520

Ginsburg, S., Eva, K., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med, 88*(10), 1539-1544. doi: 10.1097/ACM.0b013e3182a36c3d

Ginsburg, S., Gold, W., Cavalcanti, R. B., Kurabi, B., & McDonald-Blumer, H. (2011). Competencies "plus": the nature of written comments on internal medicine residents' evaluation forms. *Acad Med, 86*(10 Suppl), S30-34. doi: 10.1097/ACM.0b013e31822a6d92

Gjeraa, K., Jepsen, R. M., Rewers, M., Ostergaard, D., & Dieckmann, P. (2016). Exploring the relationship between anaesthesiologists' non-technical and technical skills. *Acta Anaesthesiol Scand, 60*(1), 36-47. doi: 10.1111/aas.12598

GMC role in education and training. (2016) (pp. 1-1). London, UK: General Medical Council.

Goddard, A. F. (2010). Planning a consultant delivered NHS. *BMJ, 341*, c6034. doi: 10.1136/bmj.c6034

Goff, B., Mandel, L., Lentz, G., Vanblaricom, A., Oelschlager, A. M., Lee, D., . . . Nielsen, P. (2005). Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol, 192*(4), 1331-1338; discussion 1338-1340. doi: 10.1016/j.ajog.2004.12.068

Gofton, W. T., Dudek, N. L., Wood, T. J., Balaa, F., & Hamstra, S. J. (2012). The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med, 87*(10), 1401-1407. doi: 10.1097/ACM.0b013e3182677805

Gohar, A., Adams, A., Gertner, E., Sackett-Lundeen, L., Heitz, R., Engle, R., . . . Bijwadia, J. (2009). Working memory capacity is decreased in sleep-deprived internal medicine residents. *J Clin Sleep Med, 5*(3), 191-197.

Goldstein, S. D., Lindeman, B., Colbert-Getz, J., Arbella, T., Dudas, R., Lidor, A., & Sacks, B. (2014). Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores. *Am J Surg, 207*(2), 231-235. doi: 10.1016/j.amjsurg.2013.10.008

Gonczi, A. (1997). Future directions for vocational education in Australian secondary schools. *Aust Nz J Voc Edu Res, 5*(1), 77-108.

Gonzalo, J. D., Heist, B. S., Duffy, B. L., Dyrbye, L., Fagan, M. J., Ferenchick, G., . . . Elnicki, M. D. (2014). Content and timing of feedback and reflection: a multi-center qualitative study of experienced bedside teachers. *BMC Med Educ, 14*, 212. doi: 10.1186/1472-6920-14-212

Gonzalo, J. D., Yang, J. J., & Huang, G. C. (2012). Systems-based content in medical morbidity and mortality conferences: a decade of change. *J Grad Med Educ, 4*(4), 438-444. doi: 10.4300/JGME-D-12-00016.1

Good, D. W., Khan, N., Kiely, E., & Brady, C. (2013). The impact of rolling theatre closures on core urology training. *Ir Med J, 106*(5), 149-151.

Good Medical Practice. (2014) (pp. 1-36). London, UK: General Medical Council.

The Good medical practice framework for appraisal and validation. (2013) (pp. 1-8). London, UK: General Medical Council.

Good Surgical Practice. (2014) (pp. 1-60). London, UK: Royal College of Surgeons of England

Goodman, C. M. (1987). The Delphi technique: a critique. *J Adv Nurs, 12*(6), 729-734.

Goova, M. T., Hollett, L. A., Tesfay, S. T., Gala, R. B., Puzziferri, N., Kehdy, F. J., & Scott, D. J. (2008). Implementation, construct validity, and benefit of a proficiency-based knot-tying and suturing curriculum. *J Surg Educ, 65*(4), 309-315. doi: 10.1016/j.jsurg.2008.04.004

Govaerts, M. J., van der Vleuten, C. P., Schuwirth, L. W., & Muijtjens, A. M. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Adv Health Sci Educ Theory Pract, 12*(2), 239-260. doi: 10.1007/s10459-006-9043-1

Graham, B., Regehr, G., & Wright, J. G. (2003). Delphi as a method to establish consensus for diagnostic criteria. *J Clin Epidemiol, 56*(12), 1150-1156.

Grantcharov, T. P., & Funch-Jensen, P. (2009). Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *Am J Surg, 197*(4), 447-449. doi: 10.1016/j.amjsurg.2008.01.024

Gray, T., & Grant, J. (2012). Good Medical Practice or CanMEDS for education? *J Clin Pathol, 65*(6), 565-567. doi: 10.1136/jclinpath-2011-200583

Green, M. L., Aagaard, E. M., Caverzagie, K. J., Chick, D. A., Holmboe, E., Kane, G., . . . Iobst, W. (2009). Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ, 1*(1), 5-20. doi: 10.4300/01.01.0003

Greenberg, C. C., Regenbogen, S. E., Studdert, D. M., Lipsitz, S. R., Rogers, S. O., Zinner, M. J., & Gawande, A. A. (2007). Patterns of communication breakdowns resulting in injury to surgical patients. *J Am Coll Surg, 204*(4), 533-540. doi: 10.1016/j.jamcollsurg.2007.01.010

Greig, P. R., Higham, H., & Vaux, E. (2015). Lack of standardisation between specialties for human factors content in postgraduate training: an analysis of specialty curricula in the UK. *BMJ Qual Saf, 24*(9), 558-560. doi: 10.1136/bmjqs-2014-003684

Grillo, H. C. (2004). Edward D. Churchill and the "rectangular" surgical residency. *Surgery, 136*(5), 947-952. doi: 10.1016/j.surg.2004.09.002

Grober, E. D., & Jewett, M. A. (2006). The concept and trajectory of "operative competence" in surgical training. *Can J Surg, 49*(4), 238-240.

Guldbrand Nielsen, D., Jensen, S. L., & O'Neill, L. (2015). Clinical assessment of transthoracic echocardiography skills: a generalizability study. *BMC Med Educ, 15*, 9. doi: 10.1186/s12909-015-0294-5

Guyatt, G. H., Oxman, A. D., Kunz, R., Vist, G. E., Falck-Ytter, Y., & Schunemann, H. J. (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ, 336*(7651), 995-998. doi: 10.1136/bmj.39490.551019.BE

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., . . . Group, G. W. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ, 336*(7650), 924-926. doi: 10.1136/bmj.39489.470347.AD

Hahn, U., Krummenauer, F., Kolbl, B., Neuhann, T., Schayan-Araghi, K., Schmickler, S., . . . Neuhann, I. (2011). Determination of valid benchmarks for outcome indicators in cataract surgery: a multicenter, prospective cohort trial. *Ophthalmology, 118*(11), 2105-2112. doi: 10.1016/j.ophtha.2011.05.011

Hall, J. C., Crebbin, W., & Ellison, A. (2004). Towards a hybrid philosophy of surgical education. *ANZ J Surg, 74*(10), 908-911. doi: 10.1111/j.1445-1433.2004.03201.x

Halpern, S. D., & Detsky, A. S. (2014). Graded autonomy in medical education--managing things that go bump in the night. *N Engl J Med, 370*(12), 1086-1089. doi: 10.1056/NEJMp1315408

Halsted, W. S. (1904). The training of the surgeon. *Johns Hopkins Hosp Bull, 15*, 267-275.

Hambleton, R. K., Swaminathan, H., Algina, J., Coulson, D.B. (1978). Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments. *Rev Educ Res, 48*(1), 1-47.

Hance, J., Aggarwal, R., Stanbridge, R., Blauth, C., Munz, Y., Darzi, A., & Pepper, J. (2005). Objective assessment of technical skills in cardiac surgery. *Eur J Cardiothorac Surg, 28*(1), 157-162. doi: 10.1016/j.ejcts.2005.03.012

Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract, 13*(5), 617-632. doi: 10.1007/s10459-007-9068-0

Harden, J. R., Crosby M. H., Davis M., & Friedman R. M. (1999). AMEE Guide No. 14: Outcome-based education: Part 5-From competency to meta-competency: a model for the specification of learning outcomes. *Med Teach, 21*(6), 546-552. doi: 10.1080/01421599978951

Hawes, R., Lehman, G. A., Hast, J., O'Connor, K. W., Crabb, D. W., Lui, A., & Christiansen, P. A. (1986). Training resident physicians in fiberoptic sigmoidoscopy. How many supervised examinations are required to achieve competence? *Am J Med, 80*(3), 465-470.

Hawkins, R. E., Welcher, C. M., Holmboe, E. S., Kirk, L. M., Norcini, J. J., Simons, K. B., & Skochelak, S. E. (2015). Implementation of competency-based medical education: are we addressing the concerns and challenges? *Med Educ, 49*(11), 1086-1102. doi: 10.1111/medu.12831

Healey, A. N., Undre, S., & Vincent, C. A. (2004). Developing observational measures of performance in surgical teams. *Qual Saf Health Care, 13 Suppl 1*, i33-40. doi: 10.1136/qhc.13.suppl_1.i33

Herbers, J. E., Jr., Noel, G. L., Cooper, G. S., Harvey, J., Pangaro, L. N., & Weaver, M. J. (1989). How accurate are faculty evaluations of clinical competence? *J Gen Intern Med, 4*(3), 202-208.

Hernandez, J. D., Bann, S. D., Munz, Y., Moorthy, K., Datta, V., Martin, S., . . . Rockall, T. (2004). Qualitative and quantitative analysis of the learning curve of a simulated surgical task on the da Vinci system. *Surg Endosc, 18*(3), 372-378. doi: 10.1007/s00464-003-9047-3

Herrera-Almario, G. E., Kirk, K., Guerrero, V. T., Jeong, K., Kim, S., & Hamad, G. G. (2016). The effect of video review of resident laparoscopic surgical skills measured by self- and external assessment. *Am J Surg, 211*(2), 315-320. doi: 10.1016/j.amjsurg.2015.05.039

Hodges, B. D. (2010). A tea-steeping or i-Doc model for medical education? *Acad Med, 85*(9 Suppl), S34-44. doi: 10.1097/ACM.0b013e3181f12f32

Hoffman, R. L., Petrosky, J. A., Eskander, M. F., Selby, L. V., & Kulaylat, A. N. (2015). Feedback fundamentals in surgical education: Tips for success. *Bull Am Coll Surg, 100*(8), 35-39.

Hogle, N. J., Widmann, W. D., Ude, A. O., Hardy, M. A., & Fowler, D. L. (2008). Does training novices to criteria and does rapid acquisition of skills on laparoscopic simulators have predictive validity or are we just playing video games? *J Surg Educ, 65*(6), 431-435. doi: 10.1016/j.jsurg.2008.05.008

Holmboe, E. S., Hawkins, R. E., & Huot, S. J. (2004). Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med, 140*(11), 874-881.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Med Teach, 32*(8), 676-682. doi: 10.3109/0142159X.2010.500704

Holmboe, E. S., Ward, D. S., Reznick, R. K., Katsufrakis, P. J., Leslie, K. M., Patel, V. L., . . . Nelson, E. A. (2011). Faculty development in assessment: the missing link in competency-based medical education. *Acad Med, 86*(4), 460-467. doi: 10.1097/ACM.0b013e31820cb2a7

Holmboe, E. S., Yamazaki, K., Edgar, L., Conforti, L., Yaghmour, N., Miller, R. S., & Hamstra, S. J. (2015). Reflections on the First 2 Years of Milestone Implementation. *J Grad Med Educ, 7*(3), 506-511. doi: 10.4300/JGME-07-03-43

Hosler, M. R., Scott, I. U., Kunselman, A. R., Wolford, K. R., Oltra, E. Z., & Murray, W. B. (2012). Impact of resident participation in cataract surgery on operative time and cost. *Ophthalmology, 119*(1), 95-98. doi: 10.1016/j.ophtha.2011.06.026

Howells, N. R., Gill, H. S., Carr, A. J., Price, A. J., & Rees, J. L. (2008). Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *J Bone Joint Surg Br, 90*(4), 494-499. doi: 10.1302/0301-620X.90B4.20414

Hsu, C. C., Sandford, B.A. (2007). The Delphi Technique: Making sense of Consensus. *Practiical Assessment, Research & Evaluation, 12*(10), 1-8.

Huang, G. C., Newman, L. R., Schwartzstein, R. M., Clardy, P. F., Feller-Kopman, D., Irish, J. T., & Smith, C. C. (2009). Procedural competence in internal medicine residents: validity of a central venous catheter insertion assessment instrument. *Acad Med, 84*(8), 1127-1134. doi: 10.1097/ACM.0b013e3181acf491

Hull, L., Arora, S., Aggarwal, R., Darzi, A., Vincent, C., & Sevdalis, N. (2012). The impact of nontechnical skills on technical performance in surgery: a systematic review. *J Am Coll Surg, 214*(2), 214-230. doi: 10.1016/j.jamcollsurg.2011.10.016

Implementing CanMEDS 2015. (2015). from http://www.royalcollege.ca/portal/page/portal/rc/canmeds/about/implementing_canmeds

Imrie, K., Frank, J. R., Ahmed, N., Gorman, L., & Harris, K. A. (2013). A new era for resident duty hours in surgery calls for greater emphasis on resident wellness. *Can J Surg, 56*(5), 295-296.

Insel, A., Carofino, B., Leger, R., Arciero, R., & Mazzocca, A. D. (2009). The development of an objective model to assess arthroscopic performance. *J Bone Joint Surg Am, 91*(9), 2287-2295. doi: 10.2106/JBJS.H.01762

Intercollegiate Specialty Examination in General Surgery. (2015) (pp. 1-4). Edinburgh, UK: Joint Committee on Intercollegiate Examinations.

Intercollegiate Surgical Curriculum Overview. (2013) *The Intercollegiate Surgical Curriculum: Educating the surgeons of the future* (pp. 1-45). London, UK: Intercollegiate Surgical Curriculum Programme

Intercollegiate surgical curriculum programme. (2013). *Royal College of Surgeons of England* Retrieved October 1, 2013, from https://http://www.iscp.ac.uk/

Intercollegiate Surgical Curriculum Programme: Overarching Assessment Blueprint 2012 (2012). from https://http://www.iscp.ac.uk/static/public/overarching_blueprint2012.pdf

Intercollegiate Surgical Curriculum Programme/Good Medical Practice Blueprint. (2012) (Vol. 3rd, pp. 1-9): The Royal College of Surgeons of England.

Iobst, W. F. (2015). Building the Plane As We Fly It. *J Grad Med Educ, 7*(2), 259-261. doi: 10.4300/JGME-D-15-00095.1

Iobst, W. F., Sherbino, J., Cate, O. T., Richardson, D. L., Dath, D., Swing, S. R., . . . Frank, J. R. (2010). Competency-based medical education in postgraduate medical education. *Med Teach, 32*(8), 651-656. doi: 10.3109/0142159X.2010.500709

Ishman, S. L., Brown, D. J., Boss, E. F., Skinner, M. L., Tunkel, D. E., Stavinoha, R., & Lin, S. Y. (2010). Development and pilot testing of an operative competency assessment tool for pediatric direct laryngoscopy and rigid bronchoscopy. *Laryngoscope, 120*(11), 2294-2300. doi: 10.1002/lary.21067

An iterative and adaptive approach to CBD implementation. (2015). from http://www.royalcollege.ca/portal/page/portal/rc/resources/cbme/implementation

Jacobsen, M. E., Andersen, M. J., Hansen, C. O., & Konge, L. (2015). Testing basic competency in knee arthroscopy using a virtual reality simulator: exploring validity and reliability. *J Bone Joint Surg Am, 97*(9), 775-781. doi: 10.2106/JBJS.N.00747

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 485-514). New Yrok: Macmillan.

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. *Med Educ, 47*(7), 650-673. doi: 10.1111/medu.12220

Jelovsek, J. E., Walters, M. D., Korn, A., Klingele, C., Zite, N., Ridgeway, B., & Barber, M. D. (2010). Establishing cutoff scores on assessments of surgical skills to determine surgical competence. *Am J Obstet Gynecol, 203*(1), 81 e81-86. doi: 10.1016/j.ajog.2010.01.073

Jenison, E. L., Gil, K. M., Lendvay, T. S., & Guy, M. S. (2012). Robotic surgical skills: acquisition, maintenance, and degradation. *JSLS, 16*(2), 218-228.

Jensen, A. R., Wright, A. S., Kim, S., Horvath, K. D., & Calhoun, K. E. (2012). Educational feedback in the operating room: a gap between resident and faculty perceptions. *Am J Surg, 204*(2), 248-255. doi: 10.1016/j.amjsurg.2011.08.019

Jeong, O., Ryu, S. Y., Choi, W. Y., Piao, Z., & Park, Y. K. (2014). Risk factors and learning curve associated with postoperative morbidity of laparoscopic total gastrectomy for gastric carcinoma. *Ann Surg Oncol, 21*(9), 2994-3001. doi: 10.1245/s10434-014-3666-x

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ, 311*(7001), 376-380.

Jones, M., Carr, A., & Montgomery, J. (2010). Specialty training places. http://careers.bmj.com/careers/advice/view-article.html?id=20001684

Kalet, A., Earp, J. A., & Kowlowitz, V. (1992). How well do faculty evaluate the interviewing skills of medical students? *J Gen Intern Med, 7*(5), 499-505.

Kane, M. (1994). Validating the performance standards associated with passing scores. . *Review of Educational Research, 64*(3), 425-461.

Karamichalis, J. M., Barach, P.R. , Nathan, M. , Henaine, R. , del Nido, P. J. , Bacha. E.A. , . (2012). Assessment of technical competency in pediatric cardiac surgery. *Progress in Pediatric Cardiology, 33*, 15-20. doi: 10.1016/j.ppedcard.2011.12.003

Karlamangla, A., Tinetti, M., Guralnik, J., Studenski, S., Wetle, T., & Reuben, D. (2007). Comorbidity in older adults: nosology of impairment, diseases, and conditions. *J Gerontol A Biol Sci Med Sci, 62*(3), 296-300.

Kassam, A., Donnon, T., & Rigby, I. (2014). Validity and reliability of an in-training evaluation report to measure the CanMEDS roles in emergency medicine residents. *CJEM, 16*(2), 144-150.

Kavic, M. S. (1998). A decade of laparoscopic cholecystectomy. *JSLS, 2*(4), 319-320.

Keith, P. B., Abrams, P., McLaughlin, J.,. (1993). Using Stakeholders in Special Education Research: How does it Influence the Research Process: Virginia State Dept. of Education, Richmond.

Kempenich, J. W., Willis, R. E., Rakosi, R., Wiersch, J., & Schenarts, P. J. (2015). How do Perceptions of Autonomy Differ in General Surgery Training Between Faculty, Senior Residents, Hospital Administrators, and the General Public? A Multi-Institutional Study. *J Surg Educ, 72*(6), e193-201. doi: 10.1016/j.jsurg.2015.06.002

Khan, M. S., Bann, S. D., Darzi, A., & Butler, P. E. (2003). Use of suturing as a measure of technical competence. *Ann Plast Surg, 50*(3), 304-308; discussion 308-309.

Kiefe, C. I., Weissman, N. W., Allison, J. J., Farmer, R., Weaver, M., & Williams, O. D. (1998). Identifying achievable benchmarks of care: concepts and methodology. *Int J Qual Health Care, 10*(5), 443-447.

Kim, H. N., Gates, E., & Lo, B. (1998). What hysterectomy [corrected] patients want to know about the roles of residents and medical students in their care. *Acad Med, 73*(3), 339-341.

Kissin, E. Y., Niu, J., Balint, P., Bong, D., Evangelisto, A., Goyal, J., . . . Kaeley, G. S. (2013). Musculoskeletal ultrasound training and competency assessment program for rheumatology fellows. *J Ultrasound Med, 32*(10), 1735-1743. doi: 10.7863/ultra.32.10.1735

Klamen, D. L., Williams, R. G., Roberts, N., & Cianciolo, A. T. (2016). Competencies, milestones, and EPAs - Are those who ignore the past condemned to repeat it? *Med Teach*, 1-7. doi: 10.3109/0142159X.2015.1132831

Kneebone, R., Nestel, D., Wetzel, C., Black, S., Jacklin, R., Aggarwal, R., . . . Darzi, A. (2006). The human face of simulation: patient-focused simulation training. *Acad Med, 81*(10), 919-924. doi: 10.1097/01.ACM.0000238323.73623.c2

Koehler, R. J., & Nicandri, G. T. (2013). Using the arthroscopic surgery skill evaluation tool as a pass-fail examination. *J Bone Joint Surg Am, 95*(23), e1871-1876. doi: 10.2106/JBJS.M.00340

Kogan, J. R., & Holmboe, E. (2013). Realizing the promise and importance of performance-based assessment. *Teach Learn Med, 25 Suppl 1*, S68-74. doi: 10.1080/10401334.2013.842912

Kohn, L. T., Corrigan, J. M. , Donaldson, M.S. . (1999). To Err Is Human: Building a Safer Health System. (Washington, DC: Institute of Medicine, National Academy of Sciences;).

Kolkman, W., Van de Put, M. A., Van den Hout, W. B., Trimbos, J. B., & Jansen, F. W. (2007). Implementation of the laparoscopic simulator in a gynecological residency curriculum. *Surg Endosc, 21*(8), 1363-1368. doi: 10.1007/s00464-006-9120-9

Konge, L., Annema, J., Clementsen, P., Minddal, V., Vilmann, P., & Ringsted, C. (2013). Using virtual-reality simulation to assess performance in endobronchial ultrasound. *Respiration, 86*(1), 59-65. doi: 10.1159/000350428

Konge, L., Clementsen, P., Larsen, K. R., Arendrup, H., Buchwald, C., & Ringsted, C. (2012). Establishing pass/fail criteria for bronchoscopy performance. *Respiration, 83*(2), 140-146. doi: 10.1159/000332333

Konopasek, L., Norcini, J., & Krupat, E. (2016). Focusing on the Formative: Building an Assessment System Aimed at Student Growth and Development. *Acad Med*. doi: 10.1097/ACM.0000000000001171

Korndorffer, J. R., Jr., Dunne, J. B., Sierra, R., Stefanidis, D., Touchard, C. L., & Scott, D. J. (2005). Simulator training for laparoscopic suturing using performance goals translates to the operating room. *J Am Coll Surg, 201*(1), 23-29. doi: 10.1016/j.jamcollsurg.2005.02.021

Korndorffer, J. R., Jr., Kasten, S. J., & Downing, S. M. (2010). A call for the utilization of consensus standards in the surgical education literature. *Am J Surg, 199*(1), 99-104. doi: 10.1016/j.amjsurg.2009.08.018

Korndorffer, J. R., Jr., Scott, D. J., Sierra, R., Brunner, W. C., Dunne, J. B., Slakey, D. P., . . . Hewitt, R. L. (2005). Developing and testing competency levels for laparoscopic skills training. *Arch Surg, 140*(1), 80-84. doi: 10.1001/archsurg.140.1.80

Kwolek, C. J., Donnelly, M. B., Sloan, D. A., Birrell, S. N., Strodel, W. E., & Schwartz, R. W. (1997). Ward evaluations: should they be abandoned? *J Surg Res, 69*(1), 1-6. doi: 10.1006/jsre.1997.5001

Laeeq, K., Bhatti, N. I., Carey, J. P., Della Santina, C. C., Limb, C. J., Niparko, J. K., . . . Francis, H. W. (2009). Pilot testing of an assessment tool for competency in mastoidectomy. *Laryngoscope, 119*(12), 2402-2410. doi: 10.1002/lary.20678

Laeeq, K., Waseem, R., Weatherly, R. A., Reh, D. D., Lin, S. Y., Lane, A. P., . . . Bhatti, N. I. (2010). In-training assessment and predictors of competency in endoscopic sinus surgery. *Laryngoscope, 120*(12), 2540-2545. doi: 10.1002/lary.21134

Lees, M. C., Merani, S., Tauh, K., & Khadaroo, R. G. (2015). Perioperative factors predicting poor outcome in elderly patients following emergency general surgery: a multivariate regression analysis. *Can J Surg, 58*(5), 312-317.

Lendvay, T. S., Brand, T. C., White, L., Kowalewski, T., Jonnadula, S., Mercer, L. D., . . . Satava, R. M. (2013). Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. *J Am Coll Surg, 216*(6), 1181-1192. doi: 10.1016/j.jamcollsurg.2013.02.012

Leung, W. C. (2002). Competency based medical training: review. *BMJ, 325*(7366), 693-696.

Lewis, F. R., & Klingensmith, M. E. (2012). Issues in general surgery residency training--2012. *Ann Surg, 256*(4), 553-559. doi: 10.1097/SLA.0b013e31826bf98c

Lieberman, M. D., Kilburn, H., Lindsey, M., & Brennan, M. F. (1995). Relation of perioperative deaths to hospital volume among patients undergoing pancreatic resection for malignancy. *Ann Surg, 222*(5), 638-645.

Lin, S. Y., Laeeq, K., Ishii, M., Kim, J., Lane, A. P., Reh, D., & Bhatti, N. I. (2009). Development and pilot-testing of a feasible, reliable, and valid operative competency assessment tool for endoscopic sinus surgery. *Am J Rhinol Allergy, 23*(3), 354-359. doi: 10.2500/ajra.2009.23.3275

Lingard, L., Espin, S., Whyte, S., Regehr, G., Baker, G. R., Reznick, R., . . . Grober, E. (2004). Communication failures in the operating room: an observational classification of recurrent types and effects. *Qual Saf Health Care, 13*(5), 330-334. doi: 10.1136/qhc.13.5.330

Littlefield, J. H., DaRosa, D. A., Anderson, K. D., Bell, R. M., Nicholas, G. G., & Wolfson, P. J. (1991). Accuracy of surgery clerkship performance raters. *Acad Med, 66*(9 Suppl), S16-18.

Livingtson, S. A., Zieky, M.J. (1982). PASSING SCORES: A manual for setting standards of perfromance on educational and occupational tests (pp. 1-73): Educational testing services

Lockley, S. W., Cronin, J. W., Evans, E. E., Cade, B. E., Lee, C. J., Landrigan, C. P., . . . Safety, G. (2004). Effect of reducing interns' weekly work hours on sleep and attentional failures. *N Engl J Med, 351*(18), 1829-1837. doi: 10.1056/NEJMoa041404

Lonergan, P. E., Mulsow, J., Tanner, W. A., Traynor, O., & Tierney, S. (2010). Residents' operative experience in general surgery programs. *Ann Surg, 251*(1), 182; author reply 182-183. doi: 10.1097/SLA.0b013e3181c77026

Louridas, M., Szasz, P., de Montbrun, S., Harris, K. A., & Grantcharov, T. P. (2016). Can We Predict Technical Aptitude?: A Systematic Review. *Ann Surg, 263*(4), 673-691. doi: 10.1097/SLA.0000000000001283

Louridas, M., Szasz, P., de Montbrun, S., Harris, K.A., Grantcharov, T.G. (2016). International assessment practices along the continuum of surgical training. *Am J Surg*. doi: 10.1016/j.amjsurg.2015.12.017

MacEwan, M. J., Dudek, N. L., Wood, T. J., & Gofton, W. T. (2016). Continued Validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the Simulated Environment. *Teach Learn Med, 28*(1), 72-79. doi: 10.1080/10401334.2015.1107483

Maciver, A., et al. (2016). Residency Survival Guide (pp. 1-63). Ottawa, ON: Canadian Association of General Surgeons.

MacMillan, P. D. (2000). Classical, Generalizability, and Multifaceted Rasch Detection of Interrater Variability in Large, Sparse Data Sets. *J Exp Educ, 68 (2)*, 167-190.

MacRae, H., Regehr, G., Leadbetter, W., & Reznick, R. K. (2000). A comprehensive examination for senior surgical residents. *Am J Surg, 179*(3), 190-193.

Magnusson, K., Osborn J. (1990). The Rise of Competency-Based Education: A Deconstructionist Analysis *J Educ Thought, 24*(1), 5-13.

Maizels, M., Yerkes, E. B., Macejko, A., Hagerty, J., Chaviano, A. H., Cheng, E. Y., . . . Kaplan, W. E. (2008). A new computer enhanced visual learning method to train urology residents in pediatric orchiopexy: a prototype for Accreditation Council for Graduate

Medical Education documentation. *J Urol, 180*(4 Suppl), 1814-1818; discussion 1818. doi: 10.1016/j.juro.2008.04.077

Malik, M., Varela DA, DV., Francis, HW,. Pandian, V,. Chien, WW,. Agrawal, Y. (2011). *Using a Virtual Reality Temporal bone Simulator to Enhance Surgical Competency in Procedural Tasks: A Pilot Study* Paper presented at the Triological Society 2011 annual meeting

Marchant, E. W. (1988). Methodological problems associated with the use of the Delphi technique - some comments. *Fire Technology, 24*(1), 59-62.

Marriott, J., Purdie, H., Crossley, J., & Beard, J. D. (2011). Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. *Br J Surg, 98*(3), 450-457. doi: 10.1002/bjs.7342

Marsden, J. S. (2006). An insider's view of the American and UK medical systems. *Br J Gen Pract, 56*(522), 60-62.

Martin, J. A., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., & Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg, 84*(2), 273-278.

Martin, M., Vashisht, B., Frezza, E., Ferone, T., Lopez, B., Pahuja, M., & Spence, R. K. (1998). Competency-based instruction in critical invasive skills improves both resident performance and patient safety. *Surgery, 124*(2), 313-317.

Martino, J. P. (1993). *Technological forecasting for decision making* (M. K. Badawy Ed. Third ed.): McGraw-Hill Engineerging and Technology Management Series.

Mashaud, L. B., Castellvi, A. O., Hollett, L. A., Hogg, D. C., Tesfay, S. T., & Scott, D. J. (2010). Two-year skill retention and certification exam performance after fundamentals of laparoscopic skills training and proficiency maintenance. *Surgery, 148*(2), 194-201. doi: 10.1016/j.surg.2010.05.012

Matsuda, K., Yoshida, Y., Kawahara, Y., Tajiri, H. . (2012). How Fast Does a Surgical Resident Learn Colonoscopy? - Analysis of Competency in Colonoscopy for Ten Surgical Residents Supervised by Expert Endoscopists (Gastroenterologists) in the High-Volume Endoscopy Center in Japan. *Gastrointest Endosc, 75*(4S), 489.

Mattar, S. G., Alseidi, A. A., Jones, D. B., Jeyarajah, D. R., Swanstrom, L. L., Aye, R. W., . . . Minter, R. M. (2013). General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Ann Surg, 258*(3), 440-449. doi: 10.1097/SLA.0b013e3182a191ca

Maxim, B. R., & Dielman, T. E. (1987). Dimensionality, internal consistency and interrater reliability of clinical performance ratings. *Med Educ, 21*(2), 130-137.

McClarty, K. L., Way, W.D., Porter, A.C., Beimers, J.N., Miles, J.A. (2013). Evidence-Based Standard Setting: Establishing a Validity Framework for Cut Scores. *Educational Researcher, 42*(2), 78-88.

McCluney, A. L., Vassiliou, M. C., Kaneva, P. A., Cao, J., Stanbridge, D. D., Feldman, L. S., & Fried, G. M. (2007). FLS simulator performance predicts intraoperative laparoscopic skill. *Surg Endosc, 21*(11), 1991-1995. doi: 10.1007/s00464-007-9451-1

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb), 22*(3), 276-282.

McKee, M., Priest, P., Ginzler, M., & Black, N. (1991). How representative are members of expert panels? *Qual Assur Health Care, 3*(2), 89-94.

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Med Educ, 43*(10), 989-992. doi: 10.1111/j.1365-2923.2009.03438.x

McLeod, S. A. (2013). Behaviorist Approach. from http://www.simplypsychology.org/behaviorism.html

McLeod, S. A. (2014). Classical Conditioning. from http://www.simplypsychology.org/classical-conditioning.htm

McLeod, S. A. (2015). Skinner- Operant Conditioning. from http://www.simplypsychology.org/operant-conditioning.html

McMains, K. C., Peel, J., Weitzel, E. K., Der-Torossian, H., & Couch, M. (2015). Perception of Shame in Otolaryngology-Head and Neck Surgery Training. *Otolaryngol Head Neck Surg, 153*(5), 786-790. doi: 10.1177/0194599815598288

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ, 6*, 42. doi: 10.1186/1472-6920-6-42

Mello, M. M., Studdert, D. M., & Brennan, T. A. (2003). The new medical malpractice crisis. *N Engl J Med, 348*(23), 2281-2284. doi: 10.1056/NEJMp030064

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Vol. 3rd edition, pp. 13). New York, NY: American Council on Education and Macmillan.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Vol. 3rd edition). New York, NY: American Council on Education and Macmillan.

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from a person's Responses and Performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Meyerson, S. L., Teitelbaum, E. N., George, B. C., Schuller, M. C., DaRosa, D. A., & Fryer, J. P. (2014). Defining the autonomy gap: when expectations do not meet reality in the operating room. *J Surg Educ, 71*(6), e64-72. doi: 10.1016/j.jsurg.2014.05.002

Milestones. (2014). *Accreditation Council for Graduate Medical Education: Next Accreditation System.* from https://http://www.acgme.org/acgmeweb/tabid/430/ProgramandInstitutionalAccreditation/NextAccreditationSystem/Milestones.aspx

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Acad Med, 65*(9 Suppl), S63-67.

Mishra, A., Catchpole, K., Dale, T., & McCulloch, P. (2008). The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surg Endosc, 22*(1), 68-73. doi: 10.1007/s00464-007-9346-1

Mishra, A., Catchpole, K., & McCulloch, P. (2009). The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Qual Saf Health Care, 18*(2), 104-108. doi: 10.1136/qshc.2007.024760

Miskovic, D., Wyles, S. M., Carter, F., Coleman, M. G., & Hanna, G. B. (2011). Development, validation and implementation of a monitoring tool for training in laparoscopic colorectal surgery in the English National Training Program. *Surg Endosc, 25*(4), 1136-1142. doi: 10.1007/s00464-010-1329-y

Mislevy, R. J. (2003). On the structure of educational assessment CSE Technical Report 597. In L. S. Steinberg, Almond, R.G. (Ed.). Los Angeles, California: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Mislevy, R. J. (2011). Evidence-Centered Design for Simulation-Based Assessment - CRESST Report 800. In N. C. f. R. o. E. University of California, Standards, and Student Testing (CRESST) (Ed.), (pp. 1-31). Los Angeles, CA: The National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mitchell, E. L., Lee, D. Y., Sevdalis, N., Partsafas, A. W., Landry, G. J., Liem, T. K., & Moneta, G. L. (2011). Evaluation of distributed practice schedules on retention of a newly acquired surgical skill: a randomized trial. *Am J Surg, 201*(1), 31-39. doi: 10.1016/j.amjsurg.2010.07.040

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ, 339*, b2535. doi: 10.1136/bmj.b2535

Mohr, J., Batalden, P., & Barach, P. (2004). Integrating patient safety into the clinical microsystem. *Qual Saf Health Care, 13 Suppl 2*, ii34-38. doi: 10.1136/qhc.13.suppl_2.ii34

Monkhouse, S. (2010). Learning in the surgical workplace: necessity not luxury. *Clin Teach, 7*(3), 167-170. doi: 10.1111/j.1743-498X.2010.00359.x

Moonen-van Loon, J. M., Overeem, K., Govaerts, M. J., Verhoeven, B. H., van der Vleuten, C. P., & Driessen, E. W. (2015). The reliability of multisource feedback in competency-based assessment programs: the effects of multiple occasions and assessor groups. *Acad Med, 90*(8), 1093-1099. doi: 10.1097/ACM.0000000000000763

Moore, A. K., Grow, D. R., Bush, R. W., & Seymour, N. E. (2008). Novices outperform experienced laparoscopists on virtual reality laparoscopy simulator. *JSLS, 12*(4), 358-362.

Moore, E. J., Price, D. L., Van Abel, K. M., & Carlson, M. L. (2014). Still under the microscope: Can a surgical aptitude test predict otolaryngology resident performance? *The Laryngoscope*, n/a-n/a. doi: 10.1002/lary.24791

Morgan, L., Pickering, S. P., Hadi, M., Robertson, E., New, S., Griffin, D., . . . McCulloch, P. (2015). A combined teamwork training and work standardisation intervention in operating theatres: controlled interrupted time series study. *BMJ Qual Saf, 24*(2), 111-119. doi: 10.1136/bmjqs-2014-003204

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *J Appl Meas, 5*(2), 189-227.

Napolitano, L. M., Savarise, M., Paramo, J. C., Soot, L. C., Todd, S. R., Gregory, J., . . . Sachdeva, A. K. (2014). Are general surgery residents ready to practice? A survey of the American College of Surgeons Board of Governors and Young Fellows Association. *J Am Coll Surg, 218*(5), 1063-1072 e1031. doi: 10.1016/j.jamcollsurg.2014.02.001

Nasca, T. J., Philibert, I., Brigham, T., & Flynn, T. C. (2012). The next GME accreditation system--rationale and benefits. *N Engl J Med, 366*(11), 1051-1056. doi: 10.1056/NEJMsr1200117

Nathens, A. B., Rivara, F. P., Jurkovich, G. J., Maier, R. V., Johansen, J. M., & Thompson, D. C. (2003). Management of the injured patient: identification of research topics for systematic review using the delphi technique. *J Trauma, 54*(3), 595-601. doi: 10.1097/01.TA.0000028044.43091.74

National Steering Committee on Resident Duty Hours. (2013) *Fatigue, risk and excellence: towards a pan-Canadian consensus on resident duty hours* Ottawa, ON: Royal College of Physicians and Surgeons of Canada

Next accreditation system (NAS) milestones. (2012). *Accreditation Council for Graduate Medical Education (ACGME).* Retrieved October 8, 2013, from http://www.acgme-nas.org/milestones.html

Nicksa, G. A., Anderson, C., Fidler, R., & Stewart, L. (2015). Innovative approach using interprofessional simulation to educate surgical residents in technical and nontechnical

skills in high-risk clinical scenarios. *JAMA Surg, 150*(3), 201-207. doi: 10.1001/jamasurg.2014.2235

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., . . . Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach, 33*(3), 206-214. doi: 10.3109/0142159X.2011.551559

Norcini, J. J. (2003). Setting standards on educational tests. *Med Educ, 37*(5), 464-469.

Norcini, J. J., Shea, JA. (1997). The Credibility and Comparability of Standards. *Applied Measurement in Education, 10*(1), 39-59.

Norman, G. R., Van der Vleuten, C. P., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Med Educ, 25*(2), 119-126.

Nugent, E., Shirilla, N., Hafeez, A., O'Riordain, D. S., Traynor, O., Harrison, A. M., & Neary, P. (2013). Development and evaluation of a simulator-based laparoscopic training program for surgical novices. *Surg Endosc, 27*(1), 214-221. doi: 10.1007/s00464-012-2423-0

Nungester, R. J., Dillon, G. F., Swanson, D. B., Orr, N. A., & Powell, R. D. (1991). Standard-setting plans for the NBME comprehensive Part I and Part II examinations. *Acad Med, 66*(8), 429-433.

O'Bryan, M. C., & Dutro, J. (2008). Impact of laparoscopic cholecystectomy on resident training: fifteen years later. *J Surg Educ, 65*(5), 346-349. doi: 10.1016/j.jsurg.2008.06.004

O'Shea, J. S. (2008). Becoming a surgeon in the early 20th century: parallels to the present. *J Surg Educ, 65*(3), 236-241. doi: 10.1016/j.jsurg.2007.12.007

Objectives of Surgical Foundations Training. (2014) (1.2 ed., pp. 1-23). Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Objectives of Training in the Specialty of General Surgery. (2010) (Vol. 1.0 pp. 1-13). Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Orzech, N., Palter, V. N., Reznick, R. K., Aggarwal, R., & Grantcharov, T. P. (2012). A comparison of 2 ex vivo training curricula for advanced laparoscopic skills: a randomized controlled trial. *Ann Surg, 255*(5), 833-839. doi: 10.1097/SLA.0b013e31824aca09

The Oxford English Dictionary. (1989) (2nd ed.). Oxford: Oxford University Press.

The Oxford English Dictionary Online, Oxford University Press. (March 2015).

Paddick, J. S., Le Rolland, P. . (2010). The Annual Review of Competence Progression (ARCP) process – a short review by the Postgraduate Medical Education and Training Board.: General Medical Council.

Paisley, A. M., Baldwin, P. J., & Paterson-Brown, S. (2001). Validity of surgical simulation for the assessment of operative skill. *Br J Surg, 88*(11), 1525-1532. doi: 10.1046/j.0007-1323.2001.01880.x

Palter, V. N., Grantcharov, T., Harvey, A., & Macrae, H. M. (2011). Ex vivo technical skills training transfers to the operating room and enhances cognitive learning: a randomized controlled trial. *Ann Surg, 253*(5), 886-889. doi: 10.1097/SLA.0b013e31821263ec

Palter, V. N., & Grantcharov, T. P. (2012). Development and validation of a comprehensive curriculum to teach an advanced minimally invasive procedure: a randomized controlled trial. *Ann Surg, 256*(1), 25-32. doi: 10.1097/SLA.0b013e318258f5aa

Palter, V. N., & Grantcharov, T. P. (2014). Individualized deliberate practice on a virtual reality simulator improves technical performance of surgical novices in the operating room: a randomized controlled trial. *Ann Surg, 259*(3), 443-448. doi: 10.1097/SLA.0000000000000254

Palter, V. N., MacRae, H. M., & Grantcharov, T. P. (2011). Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: a Delphi methodology. *Am J Surg, 201*(2), 251-259. doi: 10.1016/j.amjsurg.2010.01.031

Palter, V. N., Orzech, N., Reznick, R. K., & Grantcharov, T. P. (2013). Validation of a structured training and assessment curriculum for technical skill acquisition in minimally invasive surgery: a randomized controlled trial. *Ann Surg, 257*(2), 224-230. doi: 10.1097/SLA.0b013e31827051cd

Pandey, V., Wolfe, J. H., Moorthy, K., Munz, Y., Jackson, M. J., & Darzi, A. W. (2006). Technical skills continue to improve beyond surgical training. *J Vasc Surg, 43*(3), 539-545. doi: 10.1016/j.jvs.2005.09.047

Papandria, D., Rhee, D., Ortega, G., Zhang, Y., Gorgy, A., Makary, M. A., & Abdullah, F. (2012). Assessing trainee impact on operative time for common general surgical procedures in ACS-NSQIP. *J Surg Educ, 69*(2), 149-155. doi: 10.1016/j.jsurg.2011.08.003

Parent, R. J., Plerhoples, T. A., Long, E. E., Zimmer, D. M., Teshome, M., Mohr, C. J., . . . Dutta, S. (2010). Early, intermediate, and late effects of a surgical skills "boot camp" on an objective structured assessment of technical skills: a randomized controlled study. *J Am Coll Surg, 210*(6), 984-989. doi: 10.1016/j.jamcollsurg.2010.03.006

Park, A., Kavic, S. M., Lee, T. H., & Heniford, B. T. (2007). Minimally invasive surgery: the evolution of fellowship. *Surgery, 142*(4), 505-511; discussion 511-503. doi: 10.1016/j.surg.2007.07.009

Park, A. E., Lee, T.H. (2011). Evolution of Minimally Invasive Surgery and Its Impact on Surgical Residency Training. In R. Matteoti, Ashley, S.W. (Ed.), *Minimally Invasive Surgical Oncology* (pp. 11-22). New York, NY: Springer Berlin Heidelberg.

Patel, R., Drover, A., Chafe, R. (2015). Pediatric faculty and residents' perspectives on In-Training Evaluation Reports (ITERs). *Can Med Educ J, 6*(2), 41-53.

Pattani, R., Wu, P. E., & Dhalla, I. A. (2014). Resident duty hours in Canada: past, present and future. *CMAJ, 186*(10), 761-765. doi: 10.1503/cmaj.131053

Patterson, F., Aitkenhead, A., Edwards, H., Flaxman, C., Shaw, R., & Rosselli, A. (2015). Analysis of the Situational Judgement Test for Selection to the Foundation Programme (pp. 1-53). Birmingham, UK: Work Psychology Group.

Pellegrini, C. A. (2006). Surgical education in the United States: navigating the white waters. *Ann Surg, 244*(3), 335-342. doi: 10.1097/01.sla.0000234800.08200.6c

Pellegrini, C. A. (2012). Surgical education in the United States 2010: developing intellectual, technical and human values. *Updates Surg, 64*(1), 1-3. doi: 10.1007/s13304-011-0113-4

Pellegrini, C. A., Warshaw, A. L., & Debas, H. T. (2004). Residency training in surgery in the 21st century: a new paradigm. *Surgery, 136*(5), 953-965. doi: 10.1016/j.surg.2004.09.001

Pena, G., Altree, M., Field, J., Sainsbury, D., Babidge, W., Hewett, P., & Maddern, G. (2015). Nontechnical skills training for the operating room: A prospective study using simulation and didactic workshop. *Surgery, 158*(1), 300-309. doi: 10.1016/j.surg.2015.02.008

Peracchia, A. (2001). Presidential Address: Surgical education in the third millennium. *Ann Surg, 234*(6), 709-712.

Pereira, E. A., & Dean, B. J. (2009). British surgeons' experiences of mandatory online workplace-based assessment. *J R Soc Med, 102*(7), 287-293. doi: 10.1258/jrsm.2009.080398

Pereira, E. A., & Dean, B. J. (2013). British surgeons' experiences of a mandatory online workplace based assessment portfolio resurveyed three years on. *J Surg Educ, 70*(1), 59-67. doi: 10.1016/j.jsurg.2012.06.019

Perrenot, C., Manuela, P., Nguyen, T., Jacques, H., . (2011). Results of a proficiency-based curriculum with the virtual reality robotic srugery simulator DV-Trainer. *The International Journal of Medical Robotics and Computer Assisted Surgery, S2*, 32–66.

Peters, J. H., Fried, G. M., Swanstrom, L. L., Soper, N. J., Sillin, L. F., Schirmer, B., . . . Committee, S. F. (2004). Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery, 135*(1), 21-27. doi: 10.1016/S0039

Pfluke, J. M., Parker, M., Stauffer, J. A., Paetau, A. A., Bowers, S. P., Asbun, H. J., & Smith, C. D. (2011). Laparoscopic surgery performed through a single incision: a systematic review of the current literature. *J Am Coll Surg, 212*(1), 113-118. doi: 10.1016/j.jamcollsurg.2010.09.008

Phillips, A. W., Madhavan, A., Bookless, L. R., & Macafee, D. A. (2015). Surgical Trainers' Experience and Perspectives on Workplace-Based Assessments. *J Surg Educ, 72*(5), 979-984. doi: 10.1016/j.jsurg.2015.03.015

Phitayakorn, R., Minehart, R. D., Hemingway, M. W., Pian-Smith, M. C., & Petrusa, E. (2015). The relationship between intraoperative teamwork and management skills in patient care. *Surgery, 158*(5), 1434-1440. doi: 10.1016/j.surg.2015.03.031

Pofahl, W. E., & Pories, W. J. (2003). Current status and future directions of geriatric general surgery. *J Am Geriatr Soc, 51*(7 Suppl), S351-354.

Polavarapu, H. V., Kulaylat, A. N., Sun, S., & Hamed, O. H. (2013). 100 years of surgical education: the past, present, and future. *Bull Am Coll Surg, 98*(7), 22-27.

Policies and Procedures for Certification and Fellowship. (2014) (pp. 1-37). Ottawa, ON: Royal College of Physicians and Surgeons of Canada

Population by sex and age group (2015). *Statistics Canada* from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo10a-eng.htm

Potts, J. R., 3rd. (2016). Assessment of Competence: The Accreditation Council for Graduate Medical Education/Residency Review Committee Perspective. *Surg Clin North Am, 96*(1), 15-24. doi: 10.1016/j.suc.2015.08.008

Preisler, L., Svendsen, M. B., Nerup, N., Svendsen, L. B., & Konge, L. (2015). Simulation-based training for colonoscopy: establishing criteria for competency. *Medicine (Baltimore), 94*(4), e440. doi: 10.1097/MD.0000000000000440

Primary and Final FRCA Examination Regulations. (2015). London, UK: The Royal College of Anaesthetists.

Proctor, D. D., Price, J., Dunn, K. A., Williamson, B. A., Fountain, R. J., & Minhas, B. S. (1998). Prospective evaluation of a teaching model to determine competency in performing flexible sigmoidoscopies. *Am J Gastroenterol, 93*(8), 1217-1221. doi: 10.1111/j.1572-0241.1998.00398.x

Pugh, D., Hamstra, S. J., Wood, T. J., Humphrey-Murto, S., Touchie, C., Yudkowsky, R., & Bordage, G. (2015). A procedural skills OSCE: assessing technical and non-technical skills of internal medicine residents. *Adv Health Sci Educ Theory Pract, 20*(1), 85-100. doi: 10.1007/s10459-014-9512-x

Puram, S. V., Kozin, E. D., Sethi, R., Alkire, B., Lee, D. J., Gray, S. T., . . . Cohen, M. (2015). Impact of resident surgeons on procedure length based on common pediatric otolaryngology cases. *Laryngoscope, 125*(4), 991-997. doi: 10.1002/lary.24912

Racz, J., Dubois, L., Katchky, A., & Wall, W. (2012). Elective and emergency abdominal surgery in patients 90 years of age or older. *Can J Surg, 55*(5), 322-328. doi: 10.1503/cjs.007611

Ramani, S., & Krackov, S. K. (2012). Twelve tips for giving feedback effectively in the clinical environment. *Med Teach, 34*(10), 787-791. doi: 10.3109/0142159X.2012.684916

Ray, J. J., Sznol, J. A., Teisch, L. F., Meizoso, J. P., Allen, C. J., Namias, N., . . . Schulman, C. I. (2016). Association Between American Board of Surgery In-Training Examination Scores and Resident Performance. *JAMA Surg, 151*(1), 26-31. doi: 10.1001/jamasurg.2015.3088

Rees, C. E., Cleland, J. A., Dennis, A., Kelly, N., Mattick, K., & Monrouxe, L. V. (2014). Supervised learning events in the foundation programme: a UK-wide narrative interview study. *BMJ Open, 4*(10), e005980. doi: 10.1136/bmjopen-2014-005980

A Reference Guide for Postgraduate Specialty Training in the UK: The Gold Guide. (2014) (5th ed., pp. 1-127). London, UK: General Medical Council.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med, 73*(9), 993-997.

Reid, C. M., Kim, D. Y., Mandel, J., Smith, A., & Bansal, V. (2014). Correlating surgical clerkship evaluations with performance on the National Board of Medical Examiners examination. *J Surg Res, 190*(1), 29-35. doi: 10.1016/j.jss.2014.02.031

Rekman, J., Gofton, W., Dudek, N., Gofton, T., & Hamstra, S. J. (2016). Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment. *Acad Med, 91*(2), 186-190. doi: 10.1097/ACM.0000000000001045

Rethans, J. J., Sturmans, F., Drop, R., van der Vleuten, C., & Hobus, P. (1991). Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ, 303*(6814), 1377-1380.

Reznick, R., Regehr, G., MacRae, H., Martin, J., & McCulloch, W. (1997). Testing technical skill via an innovative "bench station" examination. *Am J Surg, 173*(3), 226-230.

Ricci, C., Casadei, R., Buscemi, S., Taffurelli, G., D'Ambra, M., Pacilio, C. A., & Minni, F. (2014). Laparoscopic distal pancreatectomy: what factors are related to the learning curve? *Surg Today*. doi: 10.1007/s00595-014-0872-x

Riem, N., Boet, S., Bould, M. D., Tavares, W., & Naik, V. N. (2012). Do technical skills correlate with non-technical skills in crisis resource management: a simulation study. *Br J Anaesth, 109*(5), 723-728. doi: 10.1093/bja/aes256

Rigberg, D., Cole, M., Hiyama, D., & McFadden, D. (2000). Surgery in the nineties. *Am Surg, 66*(9), 813-816.

Ritter, E. M., McClusky, D. A., 3rd, Gallagher, A. G., Enochsson, L., & Smith, C. D. (2006). Perceptual, visuospatial, and psychomotor abilities correlate with duration of training required on a virtual-reality flexible endoscopy simulator. *Am J Surg, 192*(3), 379-384. doi: 10.1016/j.amjsurg.2006.03.003

Rivard, J. D., Vergis, A. S., Unger, B. J., Gillman, L. M., Hardy, K. M., & Park, J. (2015). The effect of blocked versus random task practice schedules on the acquisition and retention of surgical skills. *Am J Surg, 209*(1), 93-100. doi: 10.1016/j.amjsurg.2014.08.038

Roberson, D. W., Kentala, E., & Forbes, P. (2005). Development and validation of an objective instrument to measure surgical performance at tonsillectomy. *Laryngoscope, 115*(12), 2127-2137. doi: 10.1097/01.mlg.0000178329.23359.30

Roberts, C., Rothnie, I., Zoanetti, N., & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ, 44*(7), 690-698. doi: 10.1111/j.1365-2923.2010.03689.x

Roberts, C., Zoanetti, N., & Rothnie, I. (2009). Validating a multiple mini-interview question bank assessing entry-level reasoning skills in candidates for graduate-entry medicine and dentistry programmes. *Med Educ, 43*(4), 350-359. doi: 10.1111/j.1365-2923.2009.03292.x

Roberts, N. K., Brenner, M. J., Williams, R. G., Kim, M. J., & Dunnington, G. L. (2012). Capturing the teachable moment: a grounded theory study of verbal teaching interactions in the operating room. *Surgery, 151*(5), 643-650. doi: 10.1016/j.surg.2011.12.011

Robertson, E., Morgan, L., New, S., Pickering, S., Hadi, M., Collins, G., . . . McCulloch, P. (2015). Quality Improvement in Surgery Combining Lean Improvement Methods with Teamwork Training: A Controlled Before-After Study. *PLoS One, 10*(9), e0138490. doi: 10.1371/journal.pone.0138490

Robertson, E. R., Hadi, M., Morgan, L. J., Pickering, S. P., Collins, G., New, S., . . . Catchpole, K. C. (2014). Oxford NOTECHS II: a modified theatre team non-technical skills scoring system. *PLoS One, 9*(3), e90320. doi: 10.1371/journal.pone.0090320

Rolfe, I., & McPherson, J. (1995). Formative assessment: how am I doing? *Lancet, 345*(8953), 837-839.

Rooney, D. M., Hungness, E. S., Darosa, D. A., & Pugh, C. M. (2012). Can skills coaches be used to assess resident performance in the skills laboratory? *Surgery, 151*(6), 796-802. doi: 10.1016/j.surg.2012.03.016

Rorbaek-Madsen, M., Dupont, G., Kristensen, K., Holm, T., Sorensen, J., & Dahger, H. (1992). General surgery in patients aged 80 years and older. *Br J Surg, 79*(11), 1216-1218.

Rosenbaum, L., & Lamas, D. (2012). Residents' duty hours--toward an empirical narrative. *N Engl J Med, 367*(21), 2044-2049. doi: 10.1056/NEJMsr1210160

Rothman, A. I., Blackmore, D. E., Dauphinee, W. D., & Reznick, R. (1997). Tests of sequential testing in two years' results of Part 2 of the Medical Council of Canada Qualifying Examination. *Acad Med, 72*(10 Suppl 1), S22-24.

Rowe, G. W., G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International Journal of Forecasting, 15*, 353-375.

Royal College Discipline Recognition. (2014). from http://www.royalcollege.ca/portal/page/portal/rc/credentials/discipline_recognition

Royal College of Physicians and Surgeons of Canada. (2014). from http://www.royalcollege.ca/portal/page/portal/rc/public

Royal College of Physicians and Surgeons of Canada: History. (2014). from http://www.royalcollege.ca/portal/page/portal/rc/about/history

Rubin, S. E., Spady, W.G. (1984). Acheiving Excellence through outcome-based instructional delivery. *Educ Leadersh, 41*(8), 37-44.

Russ, S., Hull, L., Rout, S., Vincent, C., Darzi, A., & Sevdalis, N. (2012). Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Ann Surg, 255*(4), 804-809. doi: 10.1097/SLA.0b013e31824a9a02

Rutkow, I. (2012). History of Surgery. In C. M. Townsend, Beauchamp, R.D., Evers, M.B., & Mattox, K.L (Ed.), *Sabiston Textbook of Surgery* (19th ed., pp. 19-35). Philadelphia, PA: Elsevier.

Sachdeva, A. K., Bell, R. H., Jr., Britt, L. D., Tarpley, J. L., Blair, P. G., & Tarpley, M. J. (2007). National efforts to reform residency education in surgery. *Acad Med, 82*(12), 1200-1210. doi: 10.1097/ACM.0b013e318159e052

SAGES Advanced Laparoscopic Workshops for Surgical Residents. (2016). from http://www.sages.org/residents_courses/advanced_courses/

Salgado, J., Grantcharov, T. P., Papasavas, P. K., Gagne, D. J., & Caushaj, P. F. (2009). Technical skills assessment as part of the selection process for a fellowship in minimally invasive surgery. *Surg Endosc, 23*(3), 641-644. doi: 10.1007/s00464-008-0033-7

Sanfey, H., Ketchum, J., Bartlett, J., Markwell, S., Meier, A. H., Williams, R., & Dunnington, G. (2010). Verification of proficiency in basic skills for postgraduate year 1 residents. *Surgery, 148*(4), 759-766; discussion 766-757. doi: 10.1016/j.surg.2010.07.018

Sarosi, G. A., Jr., Silver, M. A., Ben-David, K., & Behrns, K. E. (2014). Training outcomes of preliminary surgical residents in a university and Veterans Affairs surgical residency. *JAMA Surg, 149*(11), 1127-1132. doi: 10.1001/jamasurg.2014.2054

Satava, R. M., Cuschieri, A., Hamdorf, J., & Metrics for Objective Assessment of Surgical Skills, W. (2003). Metrics for objective Assessment. *Surg Endosc, 17*(2), 220-226. doi: 10.1007/s00464-002-8869-8

Scally, C. P., Sandhu, G., Magas, C., Gauger, P. G., & Minter, R. M. (2015). Investigating the Impact of the 2011 ACGME Resident Duty Hour Regulations on Surgical Residency Programs: The Program Director Perspective. *J Am Coll Surg, 221*(4), 883-889 e881. doi: 10.1016/j.jamcollsurg.2015.07.011

Schindler, N., Corcoran, J., & DaRosa, D. (2007). Description and impact of using a standard-setting method for determining pass/fail scores in a surgery clerkship. *Am J Surg, 193*(2), 252-257. doi: 10.1016/j.amjsurg.2006.07.017

Schumacher, D. J., Englander, R., & Carraccio, C. (2013). Developing the master learner: applying learning theory to the learner, the teacher, and the learning environment. *Acad Med, 88*(11), 1635-1645. doi: 10.1097/ACM.0b013e3182a6e8f8

Schuwirth, L. W., & Van der Vleuten, C. P. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach, 33*(6), 478-485. doi: 10.3109/0142159X.2011.565828

Scott, D. J., Goova, M. T., & Tesfay, S. T. (2007). A cost-effective proficiency-based knot-tying and suturing curriculum for residency programs. *J Surg Res, 141*(1), 7-15. doi: 10.1016/j.jss.2007.02.043

Scott, D. J., Valentine, R. J., Bergen, P. C., Rege, R. V., Laycock, R., Tesfay, S. T., & Jones, D. B. (2000). Evaluating surgical competency with the American Board of Surgery In-Training Examination, skill testing, and intraoperative assessment. *Surgery, 128*(4), 613-622. doi: 10.1067/msy.2000.108115

Scott, J., Revera Morales, D., McRitchie, A., Riviello, R., Smink, D., & Yule, S. (2016). Non-technical skills and health care provision in low- and middle-income countries: a systematic review. *Med Educ, 50*(4), 441-455. doi: 10.1111/medu.12939

Sedlack, R. E. (2011). Training to competency in colonoscopy: assessing and defining competency standards. *Gastrointest Endosc, 74*(2), 355-366 e351-352. doi: 10.1016/j.gie.2011.02.019

Sedlack, R. E., Coyle, W. J., & Group, A. C. E. R. (2016). Assessment of competency in endoscopy: establishing and validating generalizable competency benchmarks for colonoscopy. *Gastrointest Endosc, 83*(3), 516-523 e511. doi: 10.1016/j.gie.2015.04.041

Self-Regulation and the Practice of Medicine. (2015). from http://www.cpso.on.ca/

Selvan, B., Cha, A., Morris, J.B., Palmerie, J., Dumon, K., Williams, N., Korus, G.,. (2011). Endoscopic skills level - Do we need to set standards and evaluation methods for a safe training and practices? . *Surg Endosc, 25S*, S241–S372.

Sevdalis, N., Lyons, M., Healey, A. N., Undre, S., Darzi, A., & Vincent, C. A. (2009). Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg, 249*(6), 1047-1051. doi: 10.1097/SLA.0b013e3181a50220

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg, 236*(4), 458-463; discussion 463-454. doi: 10.1097/01.SLA.0000028969.51489.B4

Shalhoub, J., Santos, C., Bussey, M., Eardley, I., & Allum, W. (2015). A Descriptive Analysis of the Use of Workplace-Based Assessments in UK Surgical Training. *J Surg Educ, 72*(5), 786-794. doi: 10.1016/j.jsurg.2015.03.019

Shalhoub, J., Vesey, A. T., & Fitzgerald, J. E. (2014). What evidence is there for the use of workplace-based assessment in surgical training? *J Surg Educ, 71*(6), 906-915. doi: 10.1016/j.jsurg.2014.03.013

Shane, M. D., Pettitt, B. J., Morgenthal, C. B., & Smith, C. D. (2008). Should surgical novices trade their retractors for joysticks? Videogame experience decreases the time needed to acquire surgical skills. *Surg Endosc, 22*(5), 1294-1297. doi: 10.1007/s00464-007-9614-0

Sherbino, J., Joshi, N., & Lin, M. (2015). JGME-ALiEM Hot Topics in Medical Education Online Journal Club: An Analysis of a Virtual Discussion About Resident Teachers. *J Grad Med Educ, 7*(3), 437-444. doi: 10.4300/JGME-D-15-00071.1

Shifflette, V., Mitchell, C., Mangram, A., & Dunn, E. (2012). Current approaches to journal club by general surgery programs within the Southwestern surgical congress. *J Surg Educ, 69*(2), 162-166. doi: 10.1016/j.jsurg.2011.08.006

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull, 86*(2), 420-428.

Sierles, F. S. (2003). How to do research with self-administered surveys. *Acad Psychiatry, 27*(2), 104-113. doi: 10.1176/appi.ap.27.2.104

Simpson, J. G., Furnace, J., Crosby, J., Cumming, A. D., Evans, P. A., Friedman Ben David, M., . . . MacPherson, S. G. (2002). The Scottish doctor--learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. *Med Teach, 24*(2), 136-143. doi: 10.1080/01421590220120713

Sirinek, K. R., Willis, R., & Schwesinger, W. H. (2016). Who Will Be Able to Perform Open Biliary Surgery in 2025? *J Am Coll Surg*. doi: 10.1016/j.jamcollsurg.2016.02.019

Skinner, K. A., Helsper, J. T., Deapen, D., Ye, W., & Sposto, R. (2003). Breast cancer: do specialists make a difference? *Ann Surg Oncol, 10*(6), 606-615.

Sloan, D., . Donnelly, M,. Drake, D,. Schwartz, R,. (1993). Faculty sensitivity in detecting medical students' clinical competence *Med Teach, 17*(3), 335-342.

Sloan, D. A., Donnelly, M. B., Schwartz, R. W., & Strodel, W. E. (1995). The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg, 222*(6), 735-742.

Smith, N., Harnett, J., & Furey, A. (2015). Evaluating the reliability of surgical assessment methods in an orthopedic residency program. *Can J Surg, 58*(5), 299-304.

Smith, R. (1998). All changed, changed utterly. British medicine will be transformed by the Bristol case. *BMJ, 316*(7149), 1917-1918.

Soper, N. J., & DaRosa, D. A. (2014). Presidential address: Engendering operative autonomy in surgical training. *Surgery, 156*(4), 745-751. doi: 10.1016/j.surg.2014.06.010

Sosa, J. A., Bowman, H. M., Tielsch, J. M., Powe, N. R., Gordon, T. A., & Udelsman, R. (1998). The importance of surgeon experience for clinical and economic outcomes from thyroidectomy. *Ann Surg, 228*(3), 320-330.

Special Program Training Requirements for the Clinician Investigator Program (CIP). (2015) (1.0 ed., pp. 1-3). Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Specialty and Subspecialty Certificates. (2016). *American Board of Medical Specialties* from http://www.abms.org/member-boards/specialty-subspecialty-certificates/

Specialty of General Surgery Defined. (2013). from http://www.absurgery.org/default.jsp?aboutsurgerydefined

Specialty Training Requirements in General Surgery. (2015) (3.0 ed., pp. 1-2). Ottawa, ON: Royal College of Physicians and Surgeons of Canada

Sroka, G., Feldman, L. S., Vassiliou, M. C., Kaneva, P. A., Fayez, R., & Fried, G. M. (2010). Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. *Am J Surg, 199*(1), 115-120. doi: 10.1016/j.amjsurg.2009.07.035

Stack, B. C., Jr., Siegel, E., Bodenner, D., & Carr, M. M. (2010). A study of resident proficiency with thyroid surgery: creation of a thyroid-specific tool. *Otolaryngol Head Neck Surg, 142*(6), 856-862. doi: 10.1016/j.otohns.2010.02.028

Stefanidis, D., Acker, C. E., & Greene, F. L. (2010). Performance goals on simulators boost resident motivation and skills laboratory attendance. *J Surg Educ, 67*(2), 66-70. doi: 10.1016/j.jsurg.2010.02.002

Stefanidis, D., Korndorffer, J. R., Jr., Sierra, R., Touchard, C., Dunne, J. B., & Scott, D. J. (2005). Skill retention following proficiency-based laparoscopic simulator training. *Surgery, 138*(2), 165-170. doi: 10.1016/j.surg.2005.06.002

Stefanidis, D., Sierra, R., Korndorffer, J. R., Jr., Dunne, J. B., Markley, S., Touchard, C. L., & Scott, D. J. (2006). Intensive continuing medical education course training on simulators results in proficiency for laparoscopic suturing. *Am J Surg, 191*(1), 23-27. doi: 10.1016/j.amjsurg.2005.06.046

Steigerwald, S. N., Park, J., Hardy, K. M., Gillman, L., & Vergis, A. S. (2015). The Fundamentals of Laparoscopic Surgery and LapVR evaluation metrics may not correlate with operative performance in a novice cohort. *Med Educ Online, 20*, 30024. doi: 10.3402/meo.v20.30024

Streiner, D. L., Norman, G.R., . (2008). Reliability. In D. L. Streiner, Norman, G.R., (Ed.), *Health Measurement Scales: A practical guide to their development and use* (4th ed.). Oxford: Oxford University Press.

Sullivan, M. C., Yeo, H., Roman, S. A., Jones, A. T., Bell, R. H., Jr., & Sosa, J. A. (2013). Discrepancies in training satisfaction and program completion among 2662 categorical and preliminary general surgery residents. *Ann Surg, 257*(6), 1174-1180. doi: 10.1097/SLA.0b013e3182718ef1

Surgical Education and Training Policies. (2013). *Royal Australasian College of Surgeons* Retrieved September, 2013, from http://www.surgeons.org/policies-publications/policies/surgical-education-and-training/

Sutton, E., Chase, S. C., Klein, R., Zhu, Y., Godinez, C., Youssef, Y., & Park, A. (2013). Development of simulator guidelines for resident assessment in flexible endoscopy. *Am Surg, 79*(1), 14-22.

Swanstrom, L. L., Fried, G. M., Hoffman, K. I., & Soper, N. J. (2006). Beta test results of a new system assessing competence in laparoscopic surgery. *J Am Coll Surg, 202*(1), 62-69. doi: 10.1016/j.jamcollsurg.2005.09.024

Swanstrom, L. L., Kurian, A., Dunst, C. M., Sharata, A., Bhayani, N., & Rieder, E. (2012). Long-term outcomes of an endoscopic myotomy for achalasia: the POEM procedure. *Ann Surg, 256*(4), 659-667. doi: 10.1097/SLA.0b013e31826b5212

Swing, S. R., Beeson, M. S., Carraccio, C., Coburn, M., Iobst, W., Selden, N. R., . . . Vydareny, K. (2013). Educational milestone development in the first 7 specialties to enter the next accreditation system. *J Grad Med Educ, 5*(1), 98-106. doi: 10.4300/JGME-05-01-33

Swing, S. R., Clyman, S. G., Holmboe, E. S., & Williams, R. G. (2009). Advancing resident assessment in graduate medical education. *J Grad Med Educ, 1*(2), 278-286. doi: 10.4300/JGME-D-09-00010.1

Syme-Grant, J., White, P. S., & McAleer, J. P. (2008). Measuring competence in endoscopic sinus surgery. *Surgeon, 6*(1), 37-44.

Szasz, P., Louridas, M., de Montbrun, S., Harris, K. A., & Grantcharov, T. P. (2016). Consensus-based training and assessment model for general surgery. *Br J Surg*. doi: 10.1002/bjs.10103

Szasz, P., Louridas, M., Harris, K. A., Aggarwal, R., & Grantcharov, T. P. (2015). Assessing Technical Competence in Surgical Trainees: A Systematic Review. *Ann Surg, 261*(6), 1046-1055. doi: 10.1097/SLA.0000000000000866

Takahashi, S. G., Waddell, A., Kennedy, M., Hodges, B. (2011). Integration and Implementation Issues in Competency-Based Training in Postgraduate Medical Education: Future of Medical Education in Canada Postgraduate (FMEC PG) Project consortium.

Taravella, M. J., Davidson, R., Erlanger, M., Guiton, G., & Gregory, D. (2011). Characterizing the learning curve in phacoemulsification. *J Cataract Refract Surg, 37*(6), 1069-1075. doi: 10.1016/j.jcrs.2010.12.054

Taravella, M. J., Davidson, R., Erlanger, M., Guiton, G., & Gregory, D. (2014). Time and cost of teaching cataract surgery. *J Cataract Refract Surg, 40*(2), 212-216. doi: 10.1016/j.jcrs.2013.07.045

Taylor, J. B., Binenbaum, G., Tapino, P., & Volpe, N. J. (2007). Microsurgical lab testing is a reliable method for assessing ophthalmology residents' surgical skills. *Br J Ophthalmol, 91*(12), 1691-1694. doi: 10.1136/bjo.2007.123083

Teitelbaum, E. N., Soper, N. J., Santos, B. F., Rooney, D. M., Patel, P., Nagle, A. P., & Hungness, E. S. (2014). A simulator-based resident curriculum for laparoscopic common bile duct exploration. *Surgery, 156*(4), 880-887, 890-883. doi: 10.1016/j.surg.2014.06.020

ten Cate, O., Hart, D., Ankel, F., Busari, J., Englander, R., Glasgow, N., . . . International Competency-Based Medical Education, C. (2016). Entrustment Decision Making in Clinical Training. *Acad Med, 91*(2), 191-198. doi: 10.1097/ACM.0000000000001044

ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med, 82*(6), 542-547. doi: 10.1097/ACM.0b013e31805559c7

Thinggaard, E., Bjerrum, F., Strandbygaard, J., Gogenur, I., & Konge, L. (2015). Validity of a cross-specialty test in basic laparoscopic techniques (TABLT). *Br J Surg, 102*(9), 1106-1113. doi: 10.1002/bjs.9857

Thomas, W. E. (2006). Teaching and assessing surgical competence. *Ann R Coll Surg Engl, 88*(5), 429-432. doi: 10.1308/003588406X116927

Thompson, J. S., & Prior, M. A. (1992). Quality assurance and morbidity and mortality conference. *J Surg Res, 52*(2), 97-100.

Thomsen, A. S., Kiilgaard, J. F., Kjaerbo, H., la Cour, M., & Konge, L. (2015). Simulation-based certification for cataract surgery. *Acta Ophthalmol, 93*(5), 416-421. doi: 10.1111/aos.12691

Todsen, T., & Ringsted, C. (2015). An Addition to the Technical Skills Assessment Toolbox. *Ann Surg*. doi: 10.1097/SLA.0000000000001377

Tolsgaard, M. G., Ringsted, C., Dreisler, E., Klemmensen, A., Loft, A., Sorensen, J. L., . . . Tabor, A. (2014). Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. *Ultrasound Obstet Gynecol, 43*(4), 437-443. doi: 10.1002/uog.13198

Tomorrow's Doctors. (2009) (pp. 1-105). London, UK: General Medical Council.

Torsney, K. M., Cocker, D. M., & Slesser, A. A. (2015). The modern surgeon and competency assessment: are the workplace-based assessments evidence-based? *World J Surg, 39*(3), 623-633. doi: 10.1007/s00268-014-2875-6

The Trainee Doctor. (2011) (pp. 1-61). London, UK: General Medical Council

Traynor, O. (2011). Surgical training in an era of reduced working hours. *Surgeon, 9 Suppl 1*, S1-2. doi: 10.1016/j.surge.2011.04.003

van Bockel, J. H., Bergqvist, D., Cairols, M., Liapis, C. D., Benedetti-Valentini, F., Pandey, V., . . . Board of Vascular Surgery of the European Union of Medical, S. (2008). Education in vascular surgery: critical issues around the globe-training and qualification in vascular surgery in Europe. *J Vasc Surg, 48*(6 Suppl), 69S-75S; discussion 75S. doi: 10.1016/j.jvs.2008.08.035

Van der Vleuten, C. P., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ, 25*(2), 110-118.

van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. *Med Educ, 39*(3), 309-317. doi: 10.1111/j.1365-2929.2005.02094.x

van der Vleuten, C. P., Schuwirth, L. W., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol, 24*(6), 703-719. doi: 10.1016/j.bpobgyn.2010.04.001

van Dongen, K. W., Ahlberg, G., Bonavina, L., Carter, F. J., Grantcharov, T. P., Hyltander, A., . . . Broeders, I. A. (2011). European consensus on a competency-based virtual reality training program for basic endoscopic surgical psychomotor skills. *Surg Endosc, 25*(1), 166-171. doi: 10.1007/s00464-010-1151-6

van Hove, P. D., Tuijthof, G. J., Verdaasdonk, E. G., Stassen, L. P., & Dankelman, J. (2010). Objective assessment of technical surgical skills. *Br J Surg, 97*(7), 972-987. doi: 10.1002/bjs.7115

Varela DA, D., Malik, MU., Pandian, V., Cummings, CW., Bhatti, NI. (2011). *Comparing performance between male and female residents in otolaryngology.* Paper presented at the Triological Society's 2011 Annual Meeting.

Vassiliou, M. C., Feldman, L. S., Andrew, C. G., Bergman, S., Leffondre, K., Stanbridge, D., & Fried, G. M. (2005). A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg, 190*(1), 107-113. doi: 10.1016/j.amjsurg.2005.04.004

Velmahos, G. C., Toutouzas, K. G., Sillin, L. F., Chan, L., Clark, R. E., Theodorou, D., & Maupin, F. (2004). Cognitive task analysis for teaching technical skills in an inanimate surgical skills laboratory. *Am J Surg, 187*(1), 114-119.

Verdaasdonk, E. G., Dankelman, J., Lange, J. F., & Stassen, L. P. (2008). Incorporation of proficiency criteria for basic laparoscopic skills training: how does it work? *Surg Endosc, 22*(12), 2609-2615. doi: 10.1007/s00464-008-9849-4

Villaneuva, T. (2010). European Working Time Directive faces challenges. *CMAJ, 182*(1), E39-40. doi: 10.1503/cmaj.109-3111

von Strauss Und Torney, M., Dell-Kuster, S., Mechera, R., Rosenthal, R., & Langer, I. (2012). The cost of surgical training: analysis of operative time for laparoscopic cholecystectomy. *Surg Endosc, 26*(9), 2579-2586. doi: 10.1007/s00464-012-2236-1

von Websky, M. W., Vitz, M., Raptis, D. A., Rosenthal, R., Clavien, P. A., & Hahnloser, D. (2012). Basic laparoscopic training using the Simbionix LAP Mentor: setting the standards in the novice group. *J Surg Educ, 69*(4), 459-467. doi: 10.1016/j.jsurg.2011.12.006

Voorhees, A. B. (2001). Creating and implementig competency-based learning models. *New Dir Instit Res, 110*, 83-95.

Wade, T. J., & Webb, T. P. (2013). Tackling technical skills competency: a surgical skills rating tool. *J Surg Res, 181*(1), 1-5. doi: 10.1016/j.jss.2012.05.052

Walz, M. K., Alesina, P. F., Wenger, F. A., Deligiannis, A., Szuczik, E., Petersenn, S., . . . Mann, K. (2006). Posterior retroperitoneoscopic adrenalectomy--results of 560 procedures in 520 patients. *Surgery, 140*(6), 943-948; discussion 948-950. doi: 10.1016/j.surg.2006.07.039

Walzak, A., Bacchus, M., Schaefer, J. P., Zarnke, K., Glow, J., Brass, C., . . . Ma, I. W. (2015). Diagnosing technical competence in six bedside procedures: comparing checklists and a global rating scale in the assessment of resident performance. *Acad Med, 90*(8), 1100-1108. doi: 10.1097/ACM.0000000000000704

Wanzel, K. R., Hamstra, S. J., Anastakis, D. J., Matsumoto, E. D., & Cusimano, M. D. (2002). Effect of visual-spatial ability on learning of spatially-complex surgical skills. *Lancet, 359*(9302), 230-231. doi: 10.1016/S0140-6736(02)07441-X

Wanzel, K. R., Ward, M., & Reznick, R. K. (2002). Teaching the surgical craft: From selection to certification. *Curr Probl Surg, 39*(6), 573-659.

Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., & Regehr, G. (2003). Resident self-assessment of operative performance. *Am J Surg, 185*(6), 521-524.

Warnock, G. L. (2012). Preparing Canadian surgeons to provide care in the 21st century. *Can J Surg, 55*(4), 219-220. doi: 10.1503/cjs.016312

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet, 357*(9260), 945-949. doi: 10.1016/S0140-6736(00)04221-5

Wayne, D. B., Barsuk, J. H., O'Leary, K. J., Fudala, M. J., & McGaghie, W. C. (2008). Mastery learning of thoracentesis skills by internal medicine residents using simulation technology and deliberate practice. *J Hosp Med, 3*(1), 48-54. doi: 10.1002/jhm.268

Webster's Third New International Dictionary (1961) (3rd ed.). Springfield, MA: Merriam-Webster.

Weissman, N. W., Allison, J. J., Kiefe, C. I., Farmer, R. M., Weaver, M. T., Williams, O. D., . . . Baker, C. S. (1999). Achievable benchmarks of care: the ABCs of benchmarking. *J Eval Clin Pract, 5*(3), 269-281.

Weitz, G., Vinzentius, C., Twesten, C., Lehnert, H., Bonnemeier, H., & Konig, I. R. (2014). Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Z Med Ausbild, 31*(4), Doc41. doi: 10.3205/zma000933

Wilasrusmee, C., Lertsithichai, P., & Kittur, D. S. (2007). Vascular anastomosis model: relation between competency in a laboratory-based model and surgical competency. *Eur J Vasc Endovasc Surg, 34*(4), 405-410. doi: 10.1016/j.ejvs.2007.05.015

Wilkinson, J. R., Crossley, J. G., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ, 42*(4), 364-373. doi: 10.1111/j.1365-2923.2008.03010.x

Wilkinson, T. J., Newble, D. I., & Frampton, C. M. (2001). Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Med Educ, 35*(11), 1043-1049.

Wilkinson, T. J., & Wade, W. B. (2007). Problems with using a supervisor's report as a form of summative assessment. *Postgrad Med J, 83*(981), 504-506. doi: 10.1136/pgmj.2007.058982

William, D. (2011). What is assessment for learning? *Stud Educ Eval, 37*, 3-14.

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med, 15*(4), 270-292. doi: 10.1207/S15328015TLM1504_11

Woehr, D. J., Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *J Occup Organ Psychol, 67*, 189-205.

Wren, S. M., & Curet, M. J. (2011). Single-port robotic cholecystectomy: results from a first human use clinical study of the new da Vinci single-site surgical platform. *Arch Surg, 146*(10), 1122-1127. doi: 10.1001/archsurg.2011.143

Wu, J. S., Siewert, B., & Boiselle, P. M. (2010). Resident evaluation and remediation: a comprehensive approach. *J Grad Med Educ, 2*(2), 242-245. doi: 10.4300/JGME-D-10-00031.1

Youngson, G. G., & Flin, R. (2010). Patient safety in surgery: non-technical aspects of safe surgical performance. *Patient Saf Surg, 4*(1), 4. doi: 10.1186/1754-9493-4-4

Yudkowsky, R., Tumuluru, S., Casey, P., Herlich, N., & Ledonne, C. (2014). A patient safety approach to setting pass/fail standards for basic procedural skills checklists. *Simul Healthc, 9*(5), 277-282. doi: 10.1097/SIH.0000000000000044

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery, 139*(2), 140-149. doi: 10.1016/j.surg.2005.06.017

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., & Rowley, D. (2006). Development of a rating system for surgeons' non-technical skills. *Med Educ, 40*(11), 1098-1104. doi: 10.1111/j.1365-2929.2006.02610.x

Yule, S., Parker, S. H., Wilkinson, J., McKinley, A., MacDonald, J., Neill, A., & McAdam, T. (2015). Coaching Non-technical Skills Improves Surgical Residents' Performance in a Simulated Operating Room. *J Surg Educ, 72*(6), 1124-1130. doi: 10.1016/j.jsurg.2015.06.012

Yule, S., Rowley, D., Flin, R., Maran, N., Youngson, G., Duncan, J., & Paterson-Brown, S. (2009). Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ J Surg, 79*(3), 154-160. doi: 10.1111/j.1445-2197.2008.04833.x

Zendejas, B., Cook, D. A., Hernandez-Irizarry, R., Huebner, M., & Farley, D. R. (2012). Mastery learning simulation-based curriculum for laparoscopic TEP inguinal hernia repair. *J Surg Educ, 69*(2), 208-214. doi: 10.1016/j.jsurg.2011.08.008

Zetlitz, E., Wearing, S. C., Nicol, A., & Hart, A. M. (2012). Objective assessment of surgical training in flexor tendon repair: the utility of a low-cost porcine model as demonstrated by a single-subject research design. *J Surg Educ, 69*(4), 504-510. doi: 10.1016/j.jsurg.2012.01.001

Zevin, B., Levy, J. S., Satava, R. M., & Grantcharov, T. P. (2012). A consensus-based framework for design, validation, and implementation of simulation-based training curricula in surgery. *J Am Coll Surg, 215*(4), 580-586 e583. doi: 10.1016/j.jamcollsurg.2012.05.035

Ziv, A., Wolpe, P. R., Small, S. D., & Glick, S. (2003). Simulation-based medical education: an ethical imperative. *Acad Med, 78*(8), 783-788.

Appendices

Appendix 1. OVID MEDLINE search strategy

Database: Ovid MEDLINE(R) Daily Update <August 06, 2013>, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present>

---

1 exp Surgical Procedures, Operative/ (2353913)

2 (surgery or surgeries or surgical or surgically).tw. (1230757)

3 exp Specialties, Surgical/ (152427)

4 neurosurg*.tw. (31189)

5 operative.tw. (176650)

6 exp Laparoscopy/ (66844)

7 Laparoscopes/ (3408)

8 laparoscop*.tw. (79579)

9 minimal* invasive.tw. (34894)

10 (minimal* adj2 surg*).tw. (9066)

11 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 (3097529)

12 exp "Internship and Residency"/ (34472)

13 (resident* or residency or residencies).tw. (122607)

14 (intern or interns or interne or internes).tw. (3781)

15 12 or 13 or 14 (138969)

16 Competency-Based Education/ (2725)

17 exp "Task Performance and Analysis"/ (26905)

18 exp Educational Measurement/ (106272)

19 measure*.tw. (2230619)

20 evaluat*.tw. (2164895)

21 assess*.tw. (1744912)

22 (criterion or criteria).tw. (393638)

23 benchmark*.tw. (17654)

24 gauge*.tw. (19462)

25 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 (5163379)

26 Clinical Competence/ (63640)

27 (competent or competenc*).tw. (78345)

28 (capability or capabilities).tw. (87455)

29 expertise.tw. (24118)

30 Motor Skills/ (18898)

31 (skill or skills).tw. (102631)

32 technique*.tw. (1057410)

33 (proficient or proficiency or proficiencies).tw. (13689)

34 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33 (1373662)

35 11 and 15 and 25 and 34 (4712)

36 remove duplicates from 35 (4473)

Appendix 2. Procedures and tasks excluded in the final consensus training model, arranged by anatomic category

Anatomical categorization of tasks and procedures	Level of trainee	Median (IQR‡) per item
1. Open and minimally invasive upper gastrointestinal tract		
Laparoscopic total gastrectomy	Sr.	**2.5 (1.0)**
Open total gastrectomy	Sr.	**4.0 (1.0)**
Laparoscopic distal gastrectomy	Sr.	**3.0 (1.0)**
Laparoscopic anti-reflux procedures	Sr.	**3.5 (1.0)**
Open anti-reflux procedures	Sr.	**4.0 (1.0)**
Laparoscopic pyloromyotomy	Sr.	**3.0 (1.0)**
Open pyloromyotomy	Sr.	**4.0 (1.75)**
Gastrojejunostomy handsewn	Jr.	**3.0 (0)**
Gastrojejunostomy stapled	Jr.	**3.0 (0.75)**
OGD	Jr.	**4.0 (0.75)**
2. Open and minimally invasive lower gastrointestinal tract		
Laparoscopic splenectomy	Sr.	**4.0 (1.0)**
Laparoscopic LAR	Sr.	**4.0 (1.0)**
Laparoscopic APR	Sr.	**3.5 (1.0)**
Open APR	Sr.	**4.0 (1.5)**
Laparoscopic mobilization of right colon	Jr.	**3.0 (1.75)**
Open mobilization of right colon	Jr.	**4.0 (1.75)**
Laparoscopic mobilization of sigmoid/ left colon	Jr.	**3.0 (1.75)**
Open mobilization of sigmoid/ left colon	Jr.	**4.0 (1.75)**
Open mobilization of rectum	Jr.	**2.0 (1.75)**
Intracorporeal anastomosis (i.e. ileocolic)	Jr.	**2.5 (1.0)**
Intracorporeal anastomosis (end-to-end stapled)	Jr.	**2.5 (1.0)**
Intracorporeal vessel ligation (i.e. ileocolic/ IMA)	Jr.	**3.0 (2.0)**
Creation of stoma	Jr.	**4.0 (0.75)**
3. Open and minimally invasive hepatopancreaticobiliary		
Placement of T-tube	Sr.	**4.0 (1.5)**
Laparoscopic drainage of liver cysts	Sr.	**2.5 (1.0)**

Open drainage of liver cysts	Sr.	**2.0 (1.0)**
Laparoscopic CBD exploration	Sr.	**2.0 (0.75)**
Laparoscopic repair of liver injuries - simple	Sr.	**2.5 (1.0)**
Laparoscopic distal pancreatectomy	Sr.	**2.0 (1.5)**
Open distal pancreatectomy	Sr.	**3.0 (1.0)**
Laparoscopic cystogastrostomy	Sr.	**2.0 (1.0)**
Open cystogastrostomy	Sr.	**4.0 (1.75)**
Hepatocojejunostomy	Sr.	**2.5 (1.0)**
Pancreaticojejunostomy	Sr.	**2.0 (1.75)**
Open mobilization of gallbladder off liver	Jr.	**4.0 (1.5)**
Open cholecystectomy - simple	Jr.	**3.5 (1.0)**
Percutaneous liver biopsy	Jr.	**2.0 (1.75)**
Open liver biopsy	Jr.	**3.0 (1.75)**

4. Open and minimally invasive endocrine

Hemithyroidectomy	Sr.	**4.0 (2.5)**
Subtotal thyroidectomy	Sr.	**4.0 (2.5)**
Total thyroidectomy	Sr.	**3.5 (1.75)**
Parathyroidectomy	Sr.	**3.0 (1.75)**
Laparoscopic adrenalectomy	Sr.	**3.0 (1.0)**
Open adrenalectomy	Sr.	**3.0 (1.0)**
Opening/closing thyroid incision	Jr.	**4.0 (1.75)**
Thyroid biopsy (FNA)	Jr.	**4.0 (2.0)**

5. Open and minimally invasive hernias

TEP repair	Sr.	**3.0 (0)**
TAPP repair	Sr.	**3.0 (0)**
Laparoscopic intra-abdominal hernia repair (i.e. Petersen's space)	Sr.	**3.0 (1.75)**
Femoral hernia repair	Jr.	**4.0 (1.0)**

6. Perianal

Sphincteroplasty	Sr.	**3.0 (1.5)**
Open rectal prolapse repair	Sr.	**3.0 (1.75)**
Perineal approach rectal prolapse repair	Sr.	**2.5 (1.0)**
Fistulotomy - simple	Jr.	**4.0 (0.75)**
Hemorrhoidectomy	Jr.	**3.0 (1.0)**
Colonoscopy	Jr.	**3.5 (1.75)**

7. Breast		
Mastectomy - radical	Sr.	**4.0 (1.75)**
Mobilization of mastectomy flaps	Jr.	**3.5 (1.0)**
Mastectomy - simple	Jr.	**3.0 (0.75)**
Breast biopsy (core)	Jr.	**4.0 (1.75)**
Lumpectomy	Jr.	**3.5 (1.0)**
8. Emergency and trauma		
Abdominal packing	Jr.	**4.0 (2.0)**
Control of external hemorrhage	Jr.	**4.0 (1.5)**
Endotracheal intubation	Jr.	**4.0 (1.75)**
9. Soft tissue		
Temporal artery biopsy	Sr.	**3.0 (1.0)**
Muscle biopsy	Sr.	**4.0 (2.75)**
Superficial inguinal lymph node dissection	Sr.	**4.0 (1.75)**
Excision of malignant skin and soft tissue lesion	Jr.	**3.0 (1.0)**

‡ IQR= Q3-Q1 (The difference between the third and first quartiles)
Sr. – senior level trainee, Jr. – junior level trainee, OGD - esophagogastroduodenoscopy, LAR – low anterior resection, APR – abdominoperineal resection, IMA – inferior mesenteric artery, CBD- common bile duct, FNA – fine needle aspiration, TEP - totally extraperitoneal, TAPP - transabdominal preperitoneal.

Appendix 3. OSATS global rating instrument (J. A. Martin et al., 1997)

	5	4	3	2	1
Respect for tissue	Consistently handled tissue appropriately with minimal damage		Handled tissue carefully but occasionally caused inadvertent damage		Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments
Time and motion	Economy of movement and maximum efficiency		Efficient time/motion but some unnecessary moves		Many unnecessary moves
Instrument handling	Fluid moves with instruments and no awkwardness		Competent use of instruments although occasionally appeared stiff or awkward		Repeatedly makes tentative or awkward moves with instruments
Knowledge of instruments	Obviously familiar with the instruments required and their names		Knew the name of most instruments and used appropriate instrument for the task		Frequently asked for the wrong instrument or used an inappropriate instrument
Use of assistance	Strategically used assistant to the best advantage at all times		Good use of assistants most of the time		Consistently placed assistants poorly or failed to use assistants
Flow of operation and forward planning	Obviously planned course of operation with effortless flow from one move to the next		Demonstrated ability for forward planning with steady progression of operative procedure		Frequently stopped operating or needed to discuss next move
Knowledge of specific procedure	Demonstrated familiarity with all aspects of the operation		Knew all important aspects of the operation		Deficient knowledge. Needed specific instruction at most operative steps

Appendix 4. OSANTS global rating instrument (Dedy, Szasz, et al., 2015)

Situational awareness				
The surgeon's preparedness for the operation (e.g. knowledge of patient history), ability to perceive and gather information from the environment (people, equipment, operative progress, events, time, blood loss, etc.), make sense of the information, and anticipate potential occurrences in the near future (events, equipment needs, etc.).				
5	**4**	**3**	**2**	**1**
Surgeon well prepared, monitors/ makes sense of his/her environment throughout the procedure, and routinely considers future events / equipment needs.		Surgeon well prepared, monitors/ makes sense of his/her environment, but may shown an occasional lack of situational awareness; may occasionally fail to consider future events / equipment needs.		Surgeon ill prepared, fails to monitor/ make sense of his/her environment, resulting in a complete loss of situational awareness; repeatedly fails to consider future events / equipment needs; encounters predictable problems.
Decision making				
The surgeon's ability to engage in the decision making process by defining a problem; generating options; choosing an option and implementing an appropriate course of action; reviewing the outcomes of a plan and changing the course of action if the plan has not led to the desired outcome.				
5	**4**	**3**	**2**	**1**
Surgeon clearly and promptly defines a problem, generates option(s), makes a decision and implements it; reviews the outcome, if ineffective changes the plan without hesitation.		Surgeon defines a problem and generates option(s), but may occasionally hesitate to do so; makes / implements decisions, but occasionally appears unsure; reviews the outcome and changes the plan if necessary, but may occasionally appear hesitant / undecided.		Surgeon fails to define a problem, or generate option(s); fails to make / implement any decisions; fails to review the outcome, or adheres to a plan even if proven ineffective.
Teamwork				
The surgeon's ability to establish a shared understanding among all members of the operating room team, (e.g. by conducting a preoperative briefing, as well as a surgical pause / time-out) and maintain a shared understanding by vocalizing new information in a timely manner; the surgeon's willingness to encourage input / criticism from other team members (e.g. by asking if any team member has a concern prior to starting the operation); and to provide support and assistance to team members.				
5	**4**	**3**	**2**	**1**
Surgeon consistently establishes and maintains shared understanding among team members throughout the operation; conducts a comprehensive briefing and surgical pause; actively encourages input / criticism from team members; volunteers to provide support / assistance if required.		Surgeon strives to establish / maintain a shared understanding among team members, but shows some deficiencies in the briefing / surgical pause, and / or occasional delays in sharing new information; accepts input / criticism from team members, but does not actively encourage it; provides assistance / support to team members if requested.		Surgeon repeatedly fails to establish / maintain shared understanding among team members; omits briefing / surgical pause; fails to share new information with the team; dismisses input / criticism from team members; fails to provide support, even if requested.

Communication

The surgeon's ability to ensure effective transfer of relevant information at all times by sending clear messages, articulating effectively and adjusting voice volume to ambient noise to ensure he/she is easily heard, addressing persons directly by name / role or establishing eye contact.

5	4	3	2	1
Surgeon communicates effectively at all times by using closed-loop communication (ensuring messages are heard and understood), sending clear and complete messages, adjusting voice volume to ambient noise, and addressing persons directly by name, or establishing eye contact.		Surgeon communicates effectively, but may occasionally send incomplete or ambiguous messages, or may occasionally fail to adjust voice volume to ambient noise and / or fail to address person directly by name, or establish eye contact resulting in occasional uncertainty regarding reception / understanding of message(s).		Surgeon fails to communicate effectively, frequently sends incomplete or ambiguous messages, fails to adjust voice volume to ambient noise, fails to address person directly by name or establish eye contact, resulting in frequent uncertainty regarding reception / understanding of messages and loss of relevant information.

Leading and Directing

The surgeon's willingness and ability to assume the role of the leader in the operating room when operating as primary surgeon or assisting junior trainees; willingness to take charge if appropriate within a situation, and ability to use authority and assertiveness when needed.

5	4	3	2	1
Surgeon consistently and clearly assumes the role of the leader while operating as the primary surgeon or assisting junior trainees, takes charge in a proactive manner when appropriate within the situation, and uses authority and assertiveness when needed.		Surgeon assumes role of the leader while operating as the primary surgeon or assisting junior trainees, but may occasionally hesitate to do so or remain passive, waiting for instructions from superior; takes charge when appropriate within the situation, but with some hesitation; may occasionally lack authority and assertiveness.		Surgeon fails to assume the role of the leader when operating as primary surgeon or assisting junior trainees, always remains passive and awaits instructions from superiors, fails to take charge even in situations when it would be appropriate, and/or completely lacks authority and assertiveness.

Professionalism

The surgeon demonstrates a commitment to the patient at all times, shows accountability, is respectful towards the patient and team members, strictly adheres to standards of care and good clinical practice and through these attitudes and behaviours is a role model for team members. The surgeon maintains the aforementioned attitudes and behaviours even during stressful situations and when under pressure.

5	4	3	2	1
Surgeon consistently committed to the care of the patient, accountable, always respectful towards team members and the patient, strictly adheres to standards of care, good clinical practice, and ethics and through these attitudes and behaviours is a role model for team members; maintains professional attitudes and behaviours even in stressful situation and/or under pressure.		Surgeon committed to the care of the patient, accountable, respectful towards team members and the patient, adheres to standards of care, good clinical practice and ethics, but occasionally "cuts corners", or shows deterioration of professional attitudes and behaviours in stressful situations and/or under pressure.		Surgeon does not appear to be committed to the care of the patient, frequently shows a lack of respect for team members and the patient, disclaims responsibility for the patient; frequently cuts corners and disregards standards, behaves unethically, or shows complete deterioration or loss of previously acceptable professional attitudes and behaviours when in a stressful situation and / or under pressure.

Managing and Coordinating				
The surgeon's ability to organize activities in the operating room in a time efficient and effective way by delegating tasks and using all available resources (people, equipment, information, etc.) to achieve goals.				
5	**4**	**3**	**2**	**1**
Surgeon organizes activities in the operating room efficiently and effectively by using all available resources (people, equipment, and information, etc.) to achieve goals (e.g. by delegating tasks appropriately)		Surgeon organizes activities in the operating room effectively, but occasionally lacks efficiency by not using all available resources (people, equipment, information, etc.) to achieve goals (e.g. occasionally fails to delegate tasks appropriately).		Surgeon fails to organize activities on the operating room efficiently and effectively, fails to use available resources (people, equipment, information, etc.) to achieve goals (e.g. fails to delegate tasks).

Copyright Acknowledgements

Permission to use our previously published manuscript in section 1.6.2

Copyright Clearance Center RightsLink®

Home | Account Info | Help | Live Chat

Wolters Kluwer

Title: Assessing Technical Competence in Surgical Trainees: A Systematic Review

Author: Peter Szasz, Marisa Louridas, Kenneth Harris, et al

Publication: Annals of Surgery

Publisher: Wolters Kluwer Health, Inc.

Date: Jan 1, 2015

Copyright © 2015, Copyright (C) 2015 Wolters Kluwer Health, Inc. All rights reserved.

Logged in as:
Peter Szasz
Account #:
3001011269
LOGOUT

This reuse is free of charge. No permission letter is needed from Wolters Kluwer Health, Lippincott Williams & Wilkins. We require that all authors always include a full acknowledgement. Example: AIDS: 13 November 2013 - Volume 27 - Issue 17 - p 2679-2689. Wolters Kluwer Health Lippincott Williams & Wilkins© No modifications will be permitted.

BACK | CLOSE WINDOW

Copyright © 2016 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement. Terms and Conditions.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Permission to use our previously published manuscript in Chapter 4

This Agreement between Peter Szasz ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	3834950635420
License date	Mar 23, 2016
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	British Journal of Surgery
Licensed Content Title	Consensus-based training and assessment model for general surgery
Licensed Content Author	P. Szasz,M. Louridas,S. de Montbrun,K. A. Harris,T. P. Grantcharov
Licensed Content Date	Mar 23, 2016
Pages	1
Type of Use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Title of your thesis / dissertation	Setting Performance Standards for Competency-Based Surgical Education
Expected completion date	Nov 2016
Expected size (number of pages)	200

Permission to use our previously published manuscript in Chapter 3

This is an Agreement between Peter Szasz ("You") and Elsevier ("Elsevier"). It consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Peter Szasz
Customer address	510 king street east unit 515 Toronto, ON M5A0E5
License number	3834220592500
License date	Mar 21, 2016
Licensed content publisher	Elsevier
Licensed content publication	The American Journal of Surgery
Licensed content title	International Assessment Practices Along The Continuum of Surgical Training
Licensed content author	Marisa Louridas,Peter Szasz,Sandra de Montbrun,Kenneth A. Harris,Teodor P. Grantcharov
Licensed content date	Available online 27 February 2016
Licensed content volume number	n/a
Licensed content issue number	n/a
Number of pages	1
Start Page	None
End Page	None
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title of your thesis/dissertation	Setting Performance Standards for Competency-Based Surgical Education

Permission to use the CanMEDS diagram in section 1.3.7

Good morning Peter:

Thank you once again for taking the time to write us regarding your request to use the CanMEDS Diagram within your PhD thesis.

As your request complies with all of the requirements, it is my pleasure to grant permission for the use of the CanMEDS Diagram as you requested in your email below.

For your convenience, I have attached a PDF of the Diagram.

We simply ask that you follow the guidelines below:

1) The material is to be used for personal, educational, non-commercial purposes only. Written permission from the Royal College is required for all other uses.

2) No changes may be made to the CanMEDS Diagram without the explicit permission of the Royal College (permission is granted to use as indicated in your email request).

3) Any and all reproductions of the CanMEDS roles/Diagram must contain acknowledgement of the Royal College. Please include the following copyright text as a footnote on each slide, page and/or handout that includes the CanMEDS Diagram "Copyright © 2015 The Royal College of Physicians and Surgeons of Canada.http://rcpsc.medical.org/canmeds. Reproduced with permission."

4) Once completed, we ask that you please send us a copy of the final document for our records (please use reference number 16-286) when returning your document.

Thank you

Caroline Clouston
Administrative Assistant - CanMEDS and Faculty Development | Adjointe administrative - CanMEDS et Perfectionnement du corps professoral
Royal College of Physicians and Surgeons of Canada | Collège royal des médecins et chirurgiens du Canada
774 Echo Drive, Ottawa, ON, Canada K1S 5N8 | 774, promenade Echo, Ottawa (ON) Canada K1S 5N8
Tel/Tél 613-730-8177 ext/poste 398 | **Toll Free/Sans frais** 1-800-668-3740 ext/poste 398

Permission to use the standard setting diagrams in section 1.9.4 and 1.9.5

Permissions

T & F Reference Number: P030116-05

3/1/2016

Peter Szasz
515 - 510 King Street E
Toronto M5A0E5
Canada
peter.szasz@utoronto.ca

Dear Mr. Szasz,

We are in receipt of your request to reproduce Figure 1, Figure 2 and Figure 5 from the following article

> Steven M. Downing, Ara Tekian & Rachel Yudkowsky (2006)
> RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education
> *Teaching and Learning in Medicine: An International Journal* 18 (1): 50-57.
> DOI: 10.1207/s15328015tlm1801_11

for use in your forthcoming dissertation

Permission is granted to reproduce all editions in print and electronic.

We will be pleased to grant you permission free of charge on the condition that:

This permission is for non-exclusive English world rights.

This permission does not cover any third party copyrighted work which may appear in the material requested.

Full acknowledgement must be included showing article title, author, and full Journal title, reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com).

Thank you very much for your interest in Taylor & Francis publications. Should you have any questions or require further assistance, please feel free to contact me directly.

Sincerely,

Mary Ann Muller
Permissions Coordinator
Telephone: 215.606.4334
E-mail: maryann.muller@taylorandfrancis.com

Permissions

T & F Reference Number: P062816-07

6/28/2016

Peter Szasz
Senior Resident, General Surgery
Surgeon Scientist Training Program
University of Toronto
peter.szasz@utoronto.ca

Dear Dr. Szasz,

We are in receipt of your request to reproduce Figure 4 from the following article

> Steven M. Downing, Ara Tekian & Rachel Yudkowsky (2006)
> RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education
> *Teaching and Learning in Medicine: An International Journal* 18 (1): 50-57.
> DOI: 10.1207/s15328015tlm1801_11

for use in your forthcoming dissertation

Permission is granted to reproduce all editions in print and electronic.

We will be pleased to grant you permission free of charge on the condition that:

This permission is for non-exclusive English world rights.

This permission does not cover any third party copyrighted work which may appear in the material requested.

Full acknowledgement must be included showing article title, author, and full Journal title, reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com).

Thank you very much for your interest in Taylor & Francis publications. Should you have any questions or require further assistance, please feel free to contact me directly.

Sincerely,

Mary Ann Muller

Permissions Coordinator

Telephone: 215.606.4334

E-mail: maryann.muller@taylorandfrancis.com