Technical Skills Assessment in Robotic Surgery: Using Patient Outcomes to Set the Standard

By

Mitchell G. Goldenberg

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

Institute of Medical Science
University of Toronto

Technical Skills Assessment in Robotic Surgery: Using Patient Outcomes to Set the Standard

Mitchell G. Goldenberg
Doctor of Philosophy
Institute of Medical Science
University of Toronto
2019

1 Abstract

Introduction: Outcome-based assessments form a central component of the new wave of educational reform in surgical residency. Evaluation of technical performance in the operating room is at the heart of the move to competency-based surgical education. Despite the numerous tools available to educators to utilize within a program of assessment, key questions remain unanswered on the application of these assessment scores, particularly around setting standards for high-stakes evaluations, and their impact on clinically meaningful outcomes. This thesis focuses on the use of technical skills assessment tools in robotic surgery, the available methods of standard setting in procedural skill, and the proposal of a novel benchmarking method that uses patients' outcomes to set the standard.

Methods: Data were prospectively collected and analyzed using descriptive and inferential statistics (both parametric and non-parametric). Research designs contained within the thesis include: two systematic reviews in accordance with PRISMA guidelines, a case-control study, two methodological manuscripts, and a prospective multicenter cohort study. Contemporary validity frameworks were used throughout to provide validity evidence for all assessment tools.

Results: Six studies comprise the thesis. Two systematic reviews collated, summarized, and appraised the existing literature on robotic-assisted surgery technical skills assessment strategies, and the methods used to create absolute standards in procedural technical skill. The case-control pilot study examined a single surgeon's experience with robotic-assisted radical prostatectomy and provided important proof of the predictive relationship between technical skill and patient outcomes. A methodology paper utilized this predictive model to describe a method of standard setting that uses the clinical outcome as the fulcrum around which the technical performance benchmark is set. A prospective multicentre study used a diverse cohort of patients and surgeons to build predictive models using multiple assessment strategies across a number of clinical outcomes. Finally, an additional methodological study took the multiple predictive models to create a final, weighted composite score that can be used to set performance standards that control for important clinicopathological factors.

Conclusion: This thesis describes the predictive relationship between technical performance in robotic surgery and patient clinical outcomes, and leverages this association to create a novel method of benchmarking surgical skill.

2 Acknowledgments

I would like to thank Dr. Teodor Grantcharov for his superb guidance and mentorship over my graduate studies. His unyielding commitment to clinical and research excellence serves as a model to which all of his mentees can aspire, and I am grateful for the opportunities he continues to provide me in the academic sphere.

I would like to additionally thank Drs. Antonio Finelli and Jason Lee, who have served as members of my program advisory committee. Their input has been invaluable in helping guide my academic aspirations. I would also like to thank Dr. Rajiv Singal for his mentorship and friendship.

Thanks to the members of the Grantcharov Lab, including Dr. Peter Szasz, Dr. Andras Fecso, Dr. James Jung, Dr. Alaina Garbens, Dr. Lauren Gordon, Dr. Sara Elkabany, Dr. Bijan Dastgheib, Dr. Bojan Macanovic, Mr Karthik Raj, and Mr. Amar Chaudhry.

Thanks to the Dr. James Rutka at the Department of Surgery, Dr. Michael Fehlings at the Surgeon Scientist Training Program, and Dr. Norman Rosenblum at the Clinician Investigator Program. Thanks as well to the Chair of the Division of Urology, Dr. Neil Fleshner for allowing me the opportunity to explore the world of academia and undertake this research during residency.

Thanks to my parents Dr. Larry Goldenberg and Dr. Paula Gordon for your encouragement and guidance. You will always be professional role models to me, and your own academic successes continue to inspire me. Thanks to my brother Adam for your love and support (and for keeping me humble!). Finally, thanks to my amazing, beautiful, and talented wife Brittany Smith for her unwavering support, love, and patience during this busy phase of my life. Also, thank you to my cat Mufasa for being a wonderful distraction.

3 Funding

This research was funded through a Strategic Alignment Grant from the Royal College of Physicians and Surgeons of Canada, a grant from the Canadian Urological Oncological Group, the Draxis Health Incorporated Surgeon Scientist Fellowship (University of Toronto), the Alan S. Tauber Graduate Student Award (University of Toronto), the Dr. Michael Jewett Graduate Award (University of Toronto), and a Queen Elizabeth II/William K. Kerr Scholarship (Ontario Graduate Scholarship).

4 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. Jason Lee (Program Advisory Committee) was involved in the conception, planning, and review of the manuscript contained in Chapters 1.5, 5, and 6. He provided mentorship and guidance, and critically reviewed this thesis.

Dr. Antonio Finelli (Program Advisory Committee) was involved in the conception, planning, and review of the manuscript contained in Chapters 5 and 6. He provided mentorship and guidance, and critically reviewed this thesis.

Dr. Alaina Garbens participated in the data acquisition, interpretation, and critical review of manuscripts in Chapters 1.6 and 5.

Dr. Peter Szasz participated in the conception, planning, drafting and review of the manuscript in Chapter 1.6.

Dr. Tyler Hauer participated in the conception, planning, drafting and review of the manuscript in Chapter 1.6.

Mr. Jethro Kwong participated in the drafting and review of the manuscript in Chapter 1.5.

Dr. Anthony Costello was involved in the conception, planning, and review of the manuscript contained in Chapter 1.5.

Mr. Anton Svendrovski aided in the statistical analysis for the manuscripts in Chapters 4, 5 and 6.

Dr. Hossein Sadaat was involved in the data acquisition (as a rater) for the manuscript in Chapter 5.

Ms. Christine Neilson aided in the library search for the manuscript in Chapter 1.6.

5 Table of Contents

6 Abbreviations

AAMS	Association of American Medical Colleges
ABMS	American Board of Medical Specialties
ACGME	Accreditation Council of Graduate Medical Education
AD	Apical Dissection
AERA	American Educational Research Association
AI	Artificial Intelligence
AMEE	Association for Medical Education in Europe
APA	American Psychological Association
ARCS	Assessment of Robotic Console Skills
AUC	Area Under the Curve
BAUS	British Association of Urological Surgeons
BMI	Body Mass Index
BSTC	Basic Skills Training Curriculum
CanMEDS	Canadian Medical Education Directives for Specialists
CBD	Case-Based Discussion
CBME	Competency-Based Medical Education
CPD	Continuing Professional Development
CREB	Clinical Research Ethics Board
CVC	Central Venous Catheter
DOPS	Direct Observations of Procedural Skills
DVC	Dorsal Venous Complex
dVSS	daVinci Surgical Simulator
EBL	Estimated Blood Loss

ELRP	Extraperitoneal Laparoscopic Radical Prostatectomy
EPA	Entrustable Professional Activity
EPIC-26	Expanded Prostate Cancer Index Composite-26
ESSQ	Endoscopic Surgical Skill Qualification
EU	European Union
EWTD	European Working Time Directive
FLS	Fundamentals of Laparoscopic Surgery
FRS	Foundations of Robotic Surgery
GEARS	Global Evaluative Assessment of Robotic Skills
GERT	Generic Error Rating Tool
GMC	General Medical Council
GOALS	Global Objective Assessment of Laparoscopic Skills
GRS	Global Rating Scale
GSP	Good Surgical Practice
HFMEA	Healthcare Failure Mode Effect Analysis
HTA	Health Technology Assessment
iAE	Intraoperative Adverse Events
ICC	Intraclass Correlation
IOM	Institute of Medicine (USA)
IPSS	International Prostate Symptom Score
IQR	Interquartile Range
IRR	Interrater Reliability
LOS	Length of Stay
LRP	Laparoscopic Radical Prostatectomy
MERSQI	Medical Education Research Study Quality Instrument
MeSH	Medical Subject Headings

Mini-CEX	Mini-Clinical Evaluation Exercise
MIS	Minimally Invasive Surgery
MISTELS	McGill Inanimate System for Training and Evaluation of Laparoscopic Skills
MMC	Morbidity and Mortality Conferences
MOC	Maintenance of Certification
MUSIC	Michigan Urological Surgery Improvement Collaborative
NAS	Next Accreditation System
NCME	National Council on Measurement in Education
NOTSS	Non-Technical Skills for Surgeons
NVB	Neurovascular Bundle
O-CAT	Ottawa Clinic Assessment Tool
O-SCORE	Ottawa Surgical Competency Operating Room Evaluation
OR	Odds Ratio
ORP	Open Radical Prostatectomy
OSATS	Objective Structured Assessment of Technical Skills
OSCE	Objective Structured Clinical Examination
PACE	Prostatectomy Assessment Competency Evaluation
PGME	Postgraduate Medical Education
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ProMIS	Pro-Minimally Invasive Surgery
PSA	Prostate-Specific Antigen
PSM	Positive Surgical Margin
QBP	Quality-Based Payment
QoL	Quality of Life

R-OSATS	Robotic-Objective Structured Assessment of Technical Skills
RACE	Robotic Anastomosis Competency Evaluation
RARP	Robotic-Assisted Radical Prostatectomy
RAS	Robotic Assisted Surgery
RCPSC	Royal College of Physicians and Surgeons of Canada
ROC	Receiver Operating Characteristic
RoSS	Robotic Surgery Simulator
RPLND	Retroperitoneal Lymph Node Dissection
SBA	Simulation-Based Assessment
SEP	Simsurgery Education Platform
SME	Continuing Medical Education
SOP	Standard Operating Procedure
SSM	Simbionix Suturing Module
SV	Seminal Vesicle
SVM	Support Vector Machines
TSC	Task-Specific Checklist
UI	User Interface
UK	United Kingdom
UVA	Urethrovesical Anastomosis
VAS	Visual Analogue Scales
VR	Virtual Reality
WBA	Workplace-Based Assessment
WHO	World Health Organization

7 List of Tables

8 List of Figures

9 List of Appendices

Appendix-1: Summary of included studies assessing technical skills in robotic surgery, from “Implementing Assessments of Robotic-Assisted Technical Skill in Urologic Education: A Systematic Review and Synthesis of the Validity Evidence”

Appendix-2: Global Evaluative Assessment of Techncial Skills, from “Surgeon Performance Predicts Early Continence after Robotic-Assisted Radical Prostatectomy”

Appendix-3: Generic Error Rating Tool, from “Surgeon Performance Predicts Early Continence after Robotic-Assisted Radical Prostatectomy”

Appendix-4: Standard Setting User Interface, from “Evidence-based benchmarking in surgical performance: leveraging the skill-outcome relationship in procedural assessment”

10 Introduction

10.1 Rationale for Competency-Based Medical Education (CBME)

10.1.1 The Evolution of CBME

In order to truly appreciate the need for educational reform in surgery, it is imperative to understand the evolution of CBME over the past century. Dr. William Halsted, a general surgeon at John's Hopkins, is credited with establishing the first structured approach to surgical training, adopting a European-style model of graded exposure to clinical work under the supervision of a master-surgeon(Halsted 1904; Pellegrini 2006; O'Shea 2008). As miraculous as this new educational intervention was, it had multiple shortcomings. By relying on intense internal competition between surgical residents, a power imbalance was created that led to an obvious dichotomy between highly skilled 'chief' residents, and potentially low-skilled surgeons who were drummed out of residency(Pellegrini 2006). This model of surgical education remained in place until the 1940's, when Dr. William D. Churchill of Massachusetts General Hospital introduced a modified approach that instead relied on teamwork and collaboration among surgical trainees to care for patients(Society & Churchill n.d.; Grillo 2004). This change led to all residents completing a program of equal length, providing better training to all as opposed to a selected few. It is upon Churchill's method that our surgical education structure has sat for 60 years, until our current alignment with CBME.

The promise of CBME has been a long time coming. Nearly 100 years ago, outcome-based training was being used to reshape the industrial and business sectors, and as early as the 1960's, this concept was being explored as a means of reforming the education of elementary and high school teachers in the United States (Houston 2016). Descriptions of outcome-based training in medicine have been present in the literature for 4 decades, dating back to McGaghie's (McGaghie 1978) description, prepared for the World Health Organization (WHO) in 1978. This proposal was a response to a changing population with new and dynamic healthcare needs, requiring a system of training that

could ensure these needs are met by future physicians. In response to this demand, the medical education community has been preparing for a transition to this model in postgraduate medical education (PGME), with the literature moving from theory development to implementation over the past 40 years(Frank et al. 2010). In Canada, this need for medical education reform was addressed in the early 1990's by the Council of Ontario Faculties of Medicine, followed soon afterwards by the Mauldsley Report (Maudsley 1996) and the first iteration of the Canadian Medical Education Directives for Specialists (CanMEDS) project, which proposed a framework for competencies of a practicing physician (Frank et al. 1996).

In its essence, CBME entails the reshaping of medical education through a focus on 4 central themes: curricular outcomes, emphasizing ability, de-emphasizing time-based training, and learner-centeredness(Frank et al. 2010). This model of training takes a structured educational approach that ensures defined sets of competencies are met by trainees prior to independent practice(Albanese et al. 2008). CBME relies on iterative and utilitarian assessments of trainee performance on all aspects of health provision to provide educators with the ability to make informed decisions about the competency and readiness for practice of a given resident(Frank et al. 2010). This rigorous assessment practice also allows for trainees to have more control over their learning, through early identification of competencies or tasks that require additional attention or practice(Carraccio et al. 2002; Ericsson 2004).

CBME will have massive implications for surgical training, from trainee selection(Louridas, Szasz, Montbrun, et al. 2016) to curricular design(Keith Francis Rourke 2016), and course assessment(Holmboe et al. 2010). Understanding the multiple factors that led to the precipice of a major paradigm shift requires the examination of the issues within the status quo that led us here.

10.1.2 Issues with the Time-Based Model of Training

Traditional surgical education has taken an immersive approach to training, with defined time-periods dedicated to learning a series of tasks that have been identified over many years as central tenets in a given field of surgery(Snell & Frank 2010). This 'tea-bag' approach assumes that given enough time in clinical training, a surgical resident will acquire the skills necessary to be a competent practitioner(Snell & Frank 2010). In most surgical specialties, this constitutes 5 years of training, with residents being exposed to the various subspecialties that make up their field of practice (i.e. trauma, general surgery, urology) (David Hodges 2010). Issues arising from this model of medical education include a growing understanding of the variable rates of trainee learning(Louridas, Szasz, de Montbrun, et al. 2016). Although anecdotally many educators would argue that over a 5 year period most trainees achieve a minimum level of ability, without an outcome-focused assessment strategy it is impossible to know this for certain(David Hodges 2010). Evidence supports the notion that as early as clerkship, it is possible to identify this phenomenon using surgical skill-based tasks(Louridas et al. 2017). Additionally, issues around the reliability and frequency of skill assessments in the time-based model, as well as the focus on high-stakes or summative assessment, limits the opportunities for learning among trainees and for early identification of those residents requiring additional training or remediation(David Hodges 2010).

10.1.3 Time Restrictions in Modern Surgical Training

The need for CBME is further driven by societal pressure to limit the duty hours of medical trainees. This demand to reduce the daily or consecutive hours residents can work is primarily intended to address rising concerns around burnout and attrition among resident physicians(Imrie et al. n.d.). In the European Union (EU), including the United Kingdom (UK), the European Working Time Directive (EWTD) has had a major impact on the amount and quality of training received by surgical residents there(B. D. Kelly et al. 2011). In 2009 , this mandate reduced the maximum time one can work in a given week

to 48 hours(Hopmans et al. 2015). Evidence suggests that this limiting of surgical trainees' working hours has led to less operative exposure for intermediate level surgical residents(Parsons et al. 2011) and a paradoxical increase in sick-leave days taken by these trainees(McIntyre et al. 2010). In the United States, a similar trend has been seen with the introduction of restricted working hours in PGME. In 2011, the Accreditation Council of Graduate Medical Education (ACGME) began to enforce a 80-hour work week, with limitations on the number of consecutive hours a medical trainee can work (Iglehart 2008). Much of the public pressure to restrict duty hours in the United States followed the infamous Libby Zion case in New York City, where a young woman tragically died in part due to the care of poorly supervised and fatigued medical trainees (Iglehart 2008). Unfortunately, American trainee's perceptions of these restrictions in working hours has been fairly negative, with survey data demonstrating that most have experienced a decrease in education, preparation for senior residency, and quality of life as a result(Drolet et al. 2013). Canada too has not been immune to these time restrictions. In 2013, a consortium of educators and residents from provincial regulators agreed on a maximum shift time of 24 hours (plus 2 hours for handover), and a maximum call schedule of one-in-four nights(Pattani et al. 2014). A handful of seminal studies in the area of duty hour restriction for surgical trainees all concluded that these restrictions do not improve either resident well-being or patient safety(Bilimoria et al. 2016; Najma Ahmed et al. 2014).

10.1.4 Gaps in Knowledge Among Graduating Residents

The traditional time-based training paradigm has led to the accreditation of physicians with wide variability in general and procedural knowledge, which has led to clear gaps in healthcare quality. In 2001, the Institute of Medicine (IOM) published their document, "Crossing the Quality Chasm", a document that addressed the changing needs of the American public in the 21st century. Included in this publication was a specific focus on improving the training and regulation of physicians, to ensure that the public's expectations around their own healthcare delivery would be met(AmericaInstitute of Medicine 2001). In 2014, the IOM published another important

report, this time exclusively focused on PGME, entitled “Graduate Medical Education that Meets the Nation’s Health Needs.” In this document, the IOM emphasizes the need to focus on innovation in medical education, and includes recommendations around improving the transparency and accountability of PGME programs in the United States. An important aspect of both reports, 13 years apart, was the substantial gap in quality that is delivered across different geopolitical regions.

10.1.5 Increasing Variability of Procedure Types

As discussed above, a key driver for change in surgical education has been the changes in operative experience and exposure among residents, in part due to changes in duty hours(Kairys et al. 2008). However, restrictions in working hours is not the sole cause of these changes, and this is important to address. Surgical training undergoing this massive shift in structure and purpose is happening amidst a larger change in the way surgical care is being delivered, on the back of huge leaps in innovation and technology(Aggarwal & Darzi 2011). Minimally invasive surgery (MIS) has become the gold standard in many operations that previously were completed using an open approach(Chung & Naveed Ahmed 2010). In the 1990’s, this shift began with the widespread introduction of laparoscopic surgery, and subsequent decreases in open surgical volume were seen across residency programs(Parsa et al. 2000). This trend continued after the turn of the century, as MIS approaches to an increasing number of surgical procedures were developed and implemented(McCoy et al. 2013). Increasingly complex approaches to common surgical procedures has also had a drastic effect on the case-mix experienced by trainees in different surgical programs, creating an even more heterogenous cohort of graduating surgeons(Eckert et al. 2010; Malangoni et al. 2013). Perhaps most concerning about this trend is that the rate of conversion to open surgery in MIS cases is not insignificant in some procedures, making it important that trainees are able to safely complete cases using a traditional open approach(Eckert et al. 2010)}. The introduction of robotic surgery has created an even wider discrepancy in training experience between residency programs, with trainees' exposure to this technology varying depending on their preceptors' expertise and comfort. This issue is

especially relevant in Canada where robotic surgery is not yet as pervasive(Mamut et al. 2011). These trends have been seen internationally as well, with recent evidence from the United Kingdom highlighting the wide variations in case experience amongst surgical trainees (Elsey et al. 2017).

10.1.6 Readiness for Practice

The overall objective of all residency programs is to adequately prepare trainees for independent practice. However, changes to the healthcare system that have necessitated educational reform have also contributed to a potential generation of surgeons who are underprepared for practice after residency(George et al. 2017) . Concerning data regarding this concept of 'readiness' have emerged from various corners of the surgical world, including general surgery and hepatobiliary surgery (Napolitano et al. 2014; Osman et al. 2015). Canadian survey data supports the idea that in the pre-CBME training model, many graduating surgical residents do not yet feel competent to complete selected core procedures(Nadler et al. 2015; Pollett & Dicks 2005; Hwang 2009). This lack of preparedness may have implications for both academic and rural surgical practices, with evidence suggesting that graduating Canadian surgeons destined for community healthcare settings are not prepared or willing to perform multiple procedure-types undertaken by current community surgeons (Gillman & Vergis 2013). Literature relaying the American experience echoes this sentiment, with a survey of US fellowship program directors demonstrating that generally, trainees commencing their fellowship training after residency were underprepared and could not safely complete a multitude of surgical procedures across the spectrum of general surgery(Mattar et al. 2013). This lack of procedural competence spanned simple technical skills like suture tying and atraumatic manipulation of tissues, as well as cognitive issues such as surgical plane recognition and lack of patient ownership. Interestingly, this lack of preparedness for fellowship and independent practice may not always be apparent to the trainees themselves, with data indicating that chief residents from American residency programs overall feel confident in their ability to safely operate independently (Friedell et al. 2014). Friedell and colleagues

support this with survey data showing that only 7% of respondents indicated they were undertaking fellowship training to improve their operating skill, with the predominant reason being an interest in a subspecialty career focus. However, other survey data questions these findings, with Coleman et al finding that nearly a quarter of graduating residents in their US survey were not confident in their own surgical skills, with this being the driver for undertaking additional fellowship training.

A particular theme worth highlighting in the discussion of preparedness for practice is surgical autonomy amongst residents. While the amount of autonomy granted to residents in the operating room is certainly in part influenced by the primary surgeon's own subjectivity(Sandhu et al. 2018), trainee factors play a role as well. Mismatches in perception between faculty and trainee competency have been suggested as a cause of the variability in resident autonomy in the operating room, with interventions created to address these gaps(Young et al. 2017). The Zwisch scale was created to try and quantify the amount of autonomy granted to surgical trainees, and the literature using this measuring tool has shown that residents are consistently given less autonomy than expected, even on 'core' procedures(Meyerson et al. 2014). Barriers to surgical autonomy in trainees has been linked to various factors, most notably a perceived lack of clinical skill, time spent with the resident, resident confidence level, and knowledge base (Teman et al. 2014). Interestingly, survey data from Teman et al indicates that when surgeons are asked about external pressures limiting resident autonomy, they cite a concern for patient outcomes as a predominant factor, despite evidence to the contrary (Siam et al. 2017). However, they also point to system-level issues such as time pressures and medicolegal risk (Teman et al. 2014).

10.1.7 The Role of Accountability

Educational reform comes at a time when transparency as a value in healthcare is at a premium. More than ever in our history, patients and the public in general are demanding to have access to more information regarding the quality of healthcare delivery, including hospital(Lindenauer et al. 2014) and physician performance (Radford

et al. 2015). Focusing on the collection and measurement of healthcare quality has only grown in importance in the Western world, with huge investments from healthcare payers and providers reflecting this concept (Casalino et al. 2016). Despite this substantial investment in quality measurement, many physicians remain sceptical regarding both the feasibility of gathering, and public interpretation of, this data (Sherman et al. 2013). Primarily, physicians are concerned regarding the potential for inadequate statistical adjustment for patient case-mix, avoidance of 'high-risk' patients, and a lack of evidence supporting the impact of publicly reporting one's patient outcomes (Werner & Asch 2005).

This revolution in healthcare quality reporting has had implications for PGME as well, with an increasing demand for medical education institutions to demonstrate competency in the training of future physicians (Baron 2013). This emphasis on ensuring high-quality education encompasses every aspect of medical training, from the individual trainee, to the learning environment, to the composition of the workforce itself (Baron 2013). It is accepted that in the current surgical training framework, the goals of curricula do not necessarily align with the needs of the public (Mellinger et al. 2015). These findings and the issues surrounding them have led to a call for an outcome-based system of education, that prioritizes concrete clinical outcomes over intermediate or surrogate measures of performance (Asch et al. 2014). Amongst other initiatives, this has led to the creation of specific instruments to help quantify the educational quality of an institution (Singh et al. 2014).

10.1.8 Inter-Program Variation

Another concern regarding existing educational approaches to medical and surgical training comes from the literature suggesting that significant variation exists between graduates of different institutions. Notable evidence indicates clearly that a lack of a standardized, evidence-based approach to curricular design has a negative impact on patient care (Asch 2009). A seminal study by Bell and colleagues examined predictors of operative experience amongst graduating general surgery residents in the

US, finding that program size (number of trainees), type (university vs. hospital vs. military-based), and geographical region are all significant predictors of surgical case exposure (Bell et al. 2009). In urology, differences in trainee experience between residency programs regarding laparoscopy (Furriel et al. 2013), urotrauma (Parker et al. 2016), infertility/andrology (Ghayda et al. 2017) and transurethral surgery (Ben-Zvi et al. 2014) have been described in the literature, both locally and internationally (Friad et al. 2014). In addition, literature from the United Kingdom demonstrates clear geographical differences in performance on qualifying examinations for surgical training (Fitzgerald & Giddings 2015). Additionally, the presence of a co-existing fellowship program may have an impact on the training experience of an institution's residents (Hanks et al. 2011; Grober et al. 2008). This discrepancy in quality between surgical residencies has led to external groups creating rankings of PGME systems, which may have long term ramifications for those low-ranked centers in the absence of validated quality indicators and medical education reform (Wilson et al. 2015).

10.1.9 Credentialing Reform

A final driver for change in medical education comes at the end of postgraduate training. Credentialing of surgeons has traditionally been a relatively unstructured endeavor, with the literature supporting this (Gurgacz et al. 2012). There has been little effort to align credentialing practices with patient outcomes (Gurgacz et al. 2012), and this forms part of a wider lack of patient-centeredness in education as described above (Mellinger et al. 2015). In the US, stakeholders are moving in the direction of outcome-based credentialing, with the Next Accreditation System (NAS) initiative (Dougherty 2013). The NAS is the culmination of the competency-based training structure, relying on assessment data across all competencies to make informed decisions on a resident's readiness for independent practice (Dougherty 2013). While this is certainly a step in the right direction, experts remain concerned about the lack of clinical outcome-focus in credentialing activities, in particular at a hospital-level (Pradarelli et al. 2015). Evidence linking surgical skill to outcome may catalyze hospitals and other providers to

develop more sophisticated methods of safely credentialing surgeons (Pradarelli et al. 2015; Tam et al. 2017).

Robotic surgery has been in the spotlight in particular, with multiple recommendation papers emerging in the literature aimed at standardizing credentialing. All stress the importance of iterative assessment and outcome-measurement, moving away from the traditional approach of short-term proctorship (Zorn et al. 2009; J. Y. Lee et al. 2011). These efforts to standardize training and credentialing in robotics is underscored in particular by disagreements regarding the volume-outcome learning curve (Schiff et al. 2013), as well as concerns regarding the increase in medicolegal cases levied against robotic surgeons (Zorn et al. 2009).

10.2 CBME Around the Globe

10.2.1 Canada

The Royal College of Physicians and Surgeons of Canada (RCPSC) has been a global leader in both the theoretical exploration, and implementation of competency-based education. In addition to creation of the widely adopted CanMEDS, updated most recently in 2015 (Frank et al. 2015), Canada was among the earliest to design and implement a competency-based curriculum at the University of Toronto (Nousiainen et al. 2018), as well as the ‘Triple C’ curriculum in Family Medicine through the College of Family Physicians of Ontario (Whitehead 2012). The Canadian approach to CMBE, and in particular the CanMEDS framework, has served as the blueprint for broad implementation of outcome-based medical training (Frank & Danoff 2007). CanMEDS categorizes competency into 7 domains, ranging from ‘Leader’ to ‘Health Advocate, with ‘Medical Expert’ as the central anchor of the framework (Frank et al. 2015). Each domain of competency in the CanMEDS framework has a series of ‘key competencies’, each made up of a number of more granular ‘enabling competencies’. Finally, each domain has a number of ‘key concepts’, which are used as an additional classification method for the enabling competencies.

Despite Canada's position as a leader in the creation of competency-based frameworks and theory, there has yet to be a widespread uptake of CBME in Canadian PGME. Currently, the plan in place includes a rolling implementation of CBME across medical and surgical specialties, with urology slated to begin in July 2018. Committees comprised of current residency program directors and medical educationalists from each specialty have met regularly at the RCPSC in the years and months leading up to the enactment of CBME, with a consensus approach used to create lists of specialty-specific competencies.

10.2.2 United States

The initial impetus for outcome-based education in the US came from the Association of American Medical Colleges (AAMC), who in 1981 first introduced the concept of all graduating physicians sharing a set of common skills, attitudes, and values (Peterson 1981). The ACGME first proposed a model of competency-training in PGME in the early 2000's as a response to what they felt was a trend toward 'overspecification' of specialty training in medical education (Batalden et al. 2002). In this initial phase, called the 'outcomes project', 6 domains of competency were introduced, closely resembling the CanMEDS framework (Batalden et al. 2002). Similar to the Canadian approach, the ACGME, in conjunction with the American Board of Medical Specialties (ABMS), allowed surgical and medical program directors and educators to craft subsets of learning objectives under each of the 6 competency domains, that can be used in the creation of specialty-specific curricula. The Outcomes Project was acknowledged as having brought about meaningful changes to residency training in the US, with improvements in resident teaching and assessment seen over the period of its implementation (Swing 2009). However, it failed to bring about the sweeping change to outcome-based accreditation practices as was intended, and was replaced in 2013 by the NAS, as described above.

The concept of ‘milestones’ as a central tenet of CBME was introduced in the United States in the late 2000’s, with Green et al’s article outlining the use of the Dreyfus and Dreyfus model of skill acquisition to establish distinct, measurable, and achievable learning objectives over the course of the internal medicine residency program (M. L. Green et al. 2009; S. E. Dreyfus & H. L. Dreyfus 1980). The use of milestones as a vehicle for achieving the objectives of CBME is being rolled out now internationally as a result of the initial work done in the US (Frank et al. 2017).

10.2.3 United Kingdom (UK)

The United Kingdom’s transition to CBME occurred in the early 2000’s as well, at the behest of regulatory bodies in England and Scotland (Ellaway et al. 2007). This decision was in part informed by the Association for Medical Education in Europe’s (AMEE) publication of a model of competency-based medical education(Smith 2009). In the United Kingdom, four domains of competency have been outlined by the General Medical Council (GMC), the primary regulatory body for physicians in the UK, and include ‘Knowledge, Skills and Performance’, ‘Safety and Quality’, ‘Communication, Partnership, and Teamwork’, and ‘Maintaining Trust’ (Great Britain & Staff 2013). These are each subdivided into standards that are not specialty-specific but are rather broadly applicable across all fields of medicine and surgery. However, the Royal College of Surgeons (England) published a 2014 set of surgery-specific standards that fall into the GMC’s competency framework, called ‘Good Surgical Practice’ (GSP) (England 2014). These standards were assembled by representatives from surgical subspecialty organizations in the UK, including The British Association of Urological Surgeons (BAUS) among others.

The UK has been a leader in the implementation of workplace-based assessment (WBA), using structured assessment practices to evaluate trainee performance in the clinical environment. While this work has been applauded by the international medical education community, recent criticism has highlighted the lack of an evidence-based approach in the implementation of this programme of assessment,

with arbitrary numbers of assessments being used as surrogates for competency (Crossley & Jolly 2012). Furthermore, educators have become disillusioned with the use of frequent and often lengthy assessment sessions, with the phrase 'box-ticking exercise' being used to emphasize the absence of true buy-in from trainees and trainers alike (Phillips et al. 2015). However, this approach is soon to be replaced by a true competency and outcome-based assessment approach, with surgical societies in particular leading this recent wave of reform (Training 2017).

10.3 Assessment Strategies in CBME

An informed assessment strategy is a key element of a competency-based surgical training curriculum (Holmboe et al. 2010). The concept of validation is of colossal importance when designing a competency-based programme of assessment, and contemporary validity frameworks allow educators to ensure that they use evaluation methods that are well-supported with evidence for their use in a given context (Cook & Hatala 2016). In addition, newly implemented surgical curricula should include assessments conducted in both the low-stakes, or formative, setting as well as the high-stakes, or summative, setting (P. S. MD et al. 2016). These two broad assessment types are distinguishable by their purpose, with formative assessments centering on skill acquisition of the learner and summative assessments focused on defensible decision-making regarding a set of competencies or progression along a training pathway (Holmboe et al. 2010). Finally, the Entrustable Professional Activities (EPA) framework has been proposed and studied as a way of structuring formative and summative assessments in CBME, to ensure that trainees achieve competency in their specialties at the time of accreditation.

10.3.1 Formative Assessment: *For* Learning

At the core of CBME lies the formative assessment, a means of ensuring that trainees remain on a course to competency while informing their learning through thoughtful feedback and deliberate practice (Holmboe et al. 2010; Ericsson 2004).

These trainee evaluations are often referred to as ‘low-stakes’, yet this should not downplay their importance in a competency-based curriculum. A crucial element of learning in medical education is direct feedback to trainees, and this can be facilitated and readily tailored by these assessment types at regular intervals (Eva et al. 2016). As the purpose of formative assessments is primary to enhance the learning of a trainee, setting a minimum standard across a group of learners is of minimal importance (Beard 2005). Rather, it is more important to identify those areas of skill that need to be improved prior to the making of a summative decision regarding that trainee’s competency (Eva et al. 2016). In the current, non-competency based training paradigm, formative assessment often is confined to unstructured, brief interactions between educators and trainees, which may lead to discrepancies in perceptions of learning amongst both parties (Jensen et al. 2012). In surgical education, these formative evaluations often take place in the context of the operating room, where regular, informal feedback on technical and cognitive skills can take place (Dougherty et al. 2013). Although tools have been designed to help facilitate these interactions intraoperatively (Davies et al. 2018; Connolly et al. 2015), a lack of educational training among surgeon-assessors (Norcini et al. 2011) and a knowledge gap regarding effective feedback strategies (Maria Ahmed et al. 2013) limits the meaningfulness of these interactions.

In the shift to a competency-based framework, a structured approach is being taken to formalize formative assessments. In order to facilitate work-based assessments (WBA) of trainee performance, two primary tools have been developed and explored across medical and surgical specialities: the Direct Observations of Procedural Skills (DOPS) and the Mini-Clinical Evaluation Exercise (Mini-CEX) (Schuwirth & Van Der Vleuten 2011b). These instruments help to form a programme of assessment that, if administered on an iterative basis, can provide stakeholders with a comprehensive and reliable framing of a trainees current level of skill or ability (Schuwirth & Van Der Vleuten 2011b; Eva et al. 2016).

10.3.2 Summative Assessment: *Of* Learning

In contrast to formative assessments, the objective of a summative assessment is to ensure that trainees have reached a minimum level of ability or skill that allows them to progress to the next stage in their learning or career (Holmboe et al. 2010). When carried out at regular intervals, these assessments allow the early identification of trainees requiring further remediation (Hawkins et al. 2015). The validity of these so called 'high-stakes' assessments is therefore reliant on the integrity of their design and purpose, and the rigour with which they are approached (P. S. MD et al. 2016). It is also essential that any decisions made regarding a trainee's competency are defensible, by creating evidence-informed standards (Downing et al. 2006). The manner in which these standards are created has been explored extensively in the literature(Cohen et al. 2013), and will be discussed in later sections of this thesis.

An important consideration when discussing summative assessment strategies is the contrast between the use of a single, high-stakes assessment, versus the amalgamation of cumulative, often lower-stakes assessment data over a period of training (Hawkins et al. 2015). Both of these approaches have been used, with the former being the predominant manner in which current (pre-CBME) summative decisions are made in surgical education (i.e. the Royal College of Physicians and Surgeons of Canada) (Canada 2018). However, the argument for using the latter approach is appealing, providing a more holistic and dynamic view of a trainee's abilities in multiple settings from the perspective of multiple assessors (Holmboe et al. 2010).

10.3.3 The Entrustability Framework in CBME

When discussing methods of assessment in CBME, one must devote attention to the Entrustable Professional Activies (EPA) framework (Cate 2005). This paradigm incorporates 'trust' as an assessment construct, and allow educators to determine when a trainee is ready for increased clinical responsibility and decreased supervision (Cate 2005). It refers to discrete professional tasks that a member of one's specialty must be

able to safely and independently perform, allowing often vague and purposefully broad ‘competencies’ be more easily observable and assessable (Cate 2013). EPAs are not meant to replace a competency framework, rather, they allow more palatable translation of competencies into the clinical world (Cate 2013). Medical and surgical specialty bodies in both Canada and the US have published consensus-formed lists of EPAs that must be satisfactorily demonstrated prior to accreditation, providing guidance for educators and program directors when shaping training curricula (Swing 2009).

There are several key elements to entrustment that have been outlined in the literature. The provision of granular and detailed descriptions of specific entrustabilities is paramount in distinguishing EPA’s from competencies (Cate et al. 2015). In the creation of an EPA-based curricula, the language used must be focused and deliberate in outlining what the expectation is of a trainee at a given point in their education. Doing so is key in allowing residents to direct their own learning, in the manner that best suits their own weaknesses (H. Peters et al. 2017). Second, EPA’s should empower trainees to function with increased responsibility in their day-to-day care of patients. The graduated levels of supervision built into many EPA assessment tools are specifically designed to create an environment where educators are able to move from direct to indirect supervision of residents(Schuwirth & Van Der Vleuten 2011a). This is particularly important in the operating room, where trainees must transition over a period of time from observers and apprentices, to independent practitioners operating without oversight (Englander et al. 2016). Finally, EPA’s rely on a better understanding of the ‘gut feeling’ that supervisors have about trainees in the clinical environment (H. Peters et al. 2017). Rather than simply accept that a given trainee is not entrustable in specific context, such as performing an operation, they should reflect on their decision in order to better uncover why this is the case. This facilitates and enhances feedback to the trainee, allowing them to be more deliberate in their training moving forward.

What may limit the implementation of EPA frameworks into surgical training? The foremost issue with using a trust-based framework to deliver a CBME curriculum in surgical education is the threat to patient safety. The operating room is a high-stakes

learning environment, one with unique challenges that have more direct consequences for patient outcome when compared to other non-surgical competencies. As such, it can be challenging for surgeon educators to graduate residents along an entrustment framework in a consistent manner. This speaks to a lack of insight into how trainee surgeon's participation in surgical activities actually translates into patient adverse events or outcomes. Despite this, current time-based training curricula are still able to produce independent practitioners at the conclusion of training, without a structured approach to entrustment or other high-stakes evaluations of technical or non-technical skill. This reflects the intrinsic ability of senior surgeons to recognize those skills that are most important to ensuring satisfactory outcomes for surgical patients, and this forms the basis of work moving forward in identifying those competencies that are most important in surgical training. Finally, this speaks to the issue of assessor selection in EPA assessments. High-stakes assessments should typically be carried out by educators with expertise in the content being tested. However, in our current adoption of entrustment frameworks in surgical training, perhaps these assessors need to have been educators for a sufficient amount of time in order to have developed this intrinsic ability to recognize when a resident is safe to move along the EPA scale of graduated supervision. Given the limitations of using only the most experienced surgeons in these entrustment decisions, more work is needed to identify and distill these metrics to enable even junior faculty or residents themselves to better understand the transition points along the entrustability continuum.

10.3.4 Frameworks for Assessment Validity

In the creation of assessment tools or even programs of assessments, validity is truly everything (Harris et al. 2017). 50 years ago, the introduction of psychometrics into medical education assessment transformed the structure and rigour with which assessment creation was approached, and allowed us to better evaluate the reliability and accuracy of assessment scores (Harris et al. 2017). Since the publication of these concepts of validation, the medical education literature has been overrun with various descriptions of assessment tool design and 'validation', both for use in clinical training

and simulation (Aydin et al. 2016; Arora et al. 2011). While it is necessary that these assessment instruments are designed and published in the context a growing CBME landscape for the purpose assessing competencies, many studies continue to be published using an outdated concept of validation (Korndorffer et al. 2010). The medical education community has correctly adopted a contemporary approach to validation theory, one that focuses on gathering evidence to support the validity of assessment scores for a discrete purpose, moving away from the idea that it is sufficient to use a 'validated' instrument in any setting(education2003 n.d.) Using this now antiquated taxonomy of validation makes it difficult to understand how best to integrate existing assessment tools into a competency-based curriculum, but work is being undertaken to use modern validity frameworks to classify these studies correctly(Cook & Hatala 2016).

10.3.4.1 Cronbach's Taxonomy

Most physicians are familiar with Cronbach's Taxonomy, which uses terms such as face validity, content validity, and construct validity. Categorizing validity evidence into "types" such as these is still pervasive in the education literature. Using language such as "we have established construct validity for this assessment" are a commonly seen example of this. This syntax is in line with 1974 guidelines created by the American Psychological Association (APA), American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME).(Association et al. 1999) According to these jointly developed guidelines, face validity refers to how well an assessment modality aligns with what it claims to measure, and studies that looked at this aspect of validation normally used questionnaires to gather evidence supporting this from participants (McDougall 2007). Criterion validity, under this outdated set of definitions, normally referred to the ability of an assessment to predict another measurement of performance, often in a different environment. Construct validity previously referred to the ability of an assessment to differentiate between participants of different levels of skill or experience(CRONBACH & MEEHL 1955).

It is important to understand why this terminology is no longer seen as acceptable in the medical education community, and what the impetus for modernizing the concept of assessment validity was. The primary reason was a shift away from whether a given assessment instrument has been 'validated' in the literature, to how that assessment's scores are interpreted in a given assessment context (Korndorffer et al. 2010; Cook & Hatala 2016). In a competency-based curriculum, it is important that assessments are implemented with a clear purpose, as described above. It makes sense that low-stakes decisions do not require the same amount of evidentiary support as a high-stakes one (Korndorffer et al. 2010). The key difference between the Cronbach Taxonomy of the 1950's and the modern approach to assessment validity is the concept that 'validity evidence' must be gathered to support the use of an assessment tool or programme in a context and learner-specific manner (Cook & Hatala 2016).

10.3.4.2 Contemporary Frameworks

10.3.4.2.1 Messick

Samuel Messick's approach to measurement validation was proposed over 50 years ago (Messick 1975), but failed to gain real traction in medical education until after the turn of the century (Cook & Beckman 2006). His structured approach to validity changed both the approach to validation of educational assessments, as well as the nomenclature used. He proposed 5 sources of validity evidence, that consolidated Cronbach's taxonomy, aside from face validity. Concerns regarding false judgments of an assessment method's validity based on physical appearance (DeVellis 2016), and a shift away from 'fidelity' to 'task alignment', caused the idea of face validity to fall out of favour (Downing & Haladyna 2004). Messick's Conceptual Framework of validity includes: Content Evidence, Response Process, Internal Structure, Relationship to Other Variables, and Consequences Evidence. Content validity refers to how well the assessment aligns with the intended underlying construct, or in simpler terms, the assessment evaluates what is supposed to (Cook & Beckman 2006). This is often provided through a clear description of how the assessment platform or rubric was

designed, often using expert consensus or participant questionnaires. Response Process evidence relates to the assessment construct and the actions, thoughts, and engagement of judges and participants (Association et al. 1999). It is often difficult to provide, but often entails a description of how the judges or assessors have been trained and/or oriented to the purpose of the evaluation, and the test security used to ensure that participants provide responses or performances that accurately reflect their true ability (Cook & Beckman 2006). Internal Structure is often used synonymously with 'reliability', that is, the consistency in which ratings or judgments are provided by those conducting the assessment (Cook & Beckman 2006). Additionally, internal structure evidence often refers to the internal consistency of a given assessment, including the performance of different participant subgroups on test-items (i.e. domains of a GRS) (Association et al. 1999). An assessment's Internal Structure validity evidence is often provided by the use of interrater reliability statistics, and other correlations as described above. Relationship to Other Variables is a more readily conceptualized source of validity evidence, referring to the correlation or association between assessment scores and expected surrogates, or measures of the same construct (Foster & Cone 1995). Finally, Consequences evidence refers to the impact that assessment scores have on the participant, their training program, the clinical environment, or society at large (Cook & Beckman 2006). This source of validity evidence is commonly overlooked, as it can be difficult to measure in the medical education literature (Brydges et al. 2015). This evidence includes the establishment of pass/fail standards in an assessment setting, as well as exploration of the impact of remediation of participants with lower test scores (Cook et al. 2014).

10.3.4.2.2 Kane

The most recently described framework of educational assessment validity comes from the mind of Michael Kane, and uses the concept of a 'validity argument' to assemble sources of evidence supporting an assessment or programme of assessment (Kane 1992). Kane moves the focus of test validation from an *itemized* to a *prioritized* approach. This allows education researchers to sequentially collect evidence to support

their validity argument, from observation of performance to the consequences of test scores (Cook et al. 2015). This argument-based approach focuses on *decisions* relating to assessment scores, and operationalizes four sources of evidence to support the use of an assessment in a given educational context: Scoring, Generalization, Extrapolation, and Implications (Kane 2013). Scoring evidence refers to the method in which assessment data is collected, from response options on a written exam, to the method of observation (live vs. video-based), to the manner in which raters are selected and trained (Cook et al. 2015). Generalization evidence supports the generalizability of test scores, and includes aspects on internal consistency and interrater reliability (Cook et al. 2015). Extrapolation evidence refers to whether the assessment is testing participants on the intended constructs, through correlations with other established measures of ability or proficiency, successful discrimination between groups of participants with different levels of experience or excellence, and the responsiveness of test scores to additional training (i.e. test-retest) (Cook et al. 2015). Finally, Implications evidence supports test validity with description of standard-setting (pass/fail) methodology, and subsequent actions based on a participant standing (i.e. remediation following failure). This includes the intended or unintended consequences of an assessment, in particular the impact of test scores on the clinical environment and outcomes (Cook et al. 2015).

10.4 Assessing Performance in the Clinical Environment

In a CBME-based medical training curriculum, assessment programmes should adequately capture and score trainee performance in the clinical environment in which they work (Lockyer et al. 2017). This allows educators and other stakeholders to be sure that a trainee has not only demonstrated competency, but has been deemed safe to carry out professional tasks in their profession, a concept embodied by the EPA framework as described above (Pugh et al. 2017). Not only are WBA a potent source of data when forming summative decisions regarding the competency of a trainee, they also provide unparalleled opportunities for formative assessment, when coupled with feedback and coaching strategies (Kogan & Holmboe 2013; Phillips et al. 2015). While

there are multiple methods of quantifying the clinical skills of trainees in a number of workplace environments, there are practical and conceptual challenges that must be addressed prior to implementing a WBA strategy in residency education (Kogan & Holmboe 2013; Cate et al. 2010).

10.4.1 Types of Workplace-Based Assessment (WBA)

Multiple types of WBA exist for use in programmes of assessment, with many using similar assessment scales to quantify trainee performance (Eardley et al. 2013). The most prevalent types of WBA's are the Direct Observation of Procedural Skills (DOPS), the Mini-Clinical Examination Exercise (Mini-CEX), and the Case-Based Discussion (CBD) (Lörwald et al. 2018). These assessment methods have predominantly been used in the UK, and unsurprisingly, most literature studying the application of these evaluations in PGME comes from this country (Lörwald et al. 2018). A DOPS assessment focuses on the ability of a trainee to carry out a procedural skill, whether at the bedside (i.e. chest tube insertion), or in the operating room. A Mini-CEX uses a clinical encounter to assess a resident's ability to efficiently and comprehensively carry out a clinical examination or history on an actual patient. Finally, a CBD is completed following a trainee's interaction with a patient during multiple aspects of their healthcare 'journey', and the evaluation focuses on the trainee's knowledge around clinical decision making and medical management of the case.

In Canada, the RCPSC has published a basic outline for the implementation of WBAs in training (W. D. N. B. G. Gofton & Bhanji 2017), and includes templates to facilitate observational assessments in CBME: EPA observations, Procedural competencies, multi-source feedback, and narrative observations.

The WBA is particularly suited for surgical residencies, where trainees are heavily involved in the provision of patient care, from the clinic, to the emergency room, to the operating room (Beard et al. 2011; Davies et al. 2018). Multiple tools have been developed to facilitate not only trainee assessment in these environments, but also

feedback and even coaching interventions. The RCPSC cites two surgery-specific tools in their outline regarding WBA implementation: the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) (W. T. Gofton et al. 2012), and the Ottawa Clinic Assessment Tool (O-CAT) (Rekman et al. 2016). These both have multiple sources of validity evidence supporting their use for the assessment of surgical trainees in the work environment, and both focus primarily on facilitating feedback and directing future learning. In promoting these two assessment scales, the RCPSC encourages educators to employ an entrustment-based approach in WBAs, citing evidence that EPA anchors are more likely to produce reliable evaluations (Crossley et al. 2011).

10.4.2 Theoretical Assumptions in WBAs

Firstly, it is important to understand underlying assumptions in WBA that should be addressed when implementing or interpreting these evaluations in the context of surgical training. Govaerts and Van Der Vleuten highlight and synthesize these as three primary suppositions (Govaerts & Van Der Vleuten 2013):

i. Learning in the workplace is a linear, discrete process;
ii. Competence is a fixed construct when abstracted from observed performance; and
iii. Performance can be objectified based on observations made by assessors.

These assumptions are being challenged by emerging literature in the fields of medical education and quality improvement (Govaerts & Van Der Vleuten 2013). Clear evidence from the field of applied psychology indicates that learning and knowledge acquisition, is a non-linear process (Deadrick 1997; Stewart & Nandkeolyar 2007; Wenghofer et al. 2009). While Miller and the Dreyfus brothers describe the acquisition of skills in a step-wise manner (Miller 1990; S. E. Dreyfus & H. L. Dreyfus 1980), so-called sociocultural theories describe a more dynamic interaction between learning and the educational context (Bleakley 2006). This interpretation of knowledge acquisition accounts for the unique social interactions that occur in the process of learning in the

workplace, and therefore argues that learning is not only non-linear, but often non-predictable.

The idea that competence is a stable, sequential construct, is often challenged in the education literature (Sturman et al. 2005). Despite this, we continue to think of variations in the performance of an individual at points in their training continuum as measurement error, and often disregard these discrepant scores when making high-stakes decisions regarding an individual's abilities (Govaerts & Van Der Vleuten 2013). However, it is essential to appreciate that this variation can be due to a host of external factors, that include testing environment (Stewart & Nandkeolyar 2007), trainee motivation or other psychological influences (Beal et al. 2005), or the performance of other members of the medical team. These all may pose threats to the use of WBA scores, if not adequately addressed or accounted for, as explained below (Govaerts & Van Der Vleuten 2013).

10.4.3 Threats to Validity and Other Challenges in WBA

As with any form of iterative, time-demanding assessment program, WBAs have been met with scrutiny during their implementation into surgical training (Bookless et al. 2015). Investigators in surgical education have yet to be fully convinced that the currently used assessment tools have sufficient validity evidence supporting their use in formative evaluations of trainee competence, let alone high-stakes assessment (Shalhoub et al. 2014). There are multiple reasons for this beyond validity data, including faculty buy in, inadequate adjustment for clinical and patient parameters, and observational sampling (van der Vleuten 1996; Govaerts & Van Der Vleuten 2013; Eardley et al. 2013; Kogan et al. 2011; Hauer et al. 2018).

Concerns surrounding a lack of supporting validity evidence for WBAs are not without merit. The Content validation of WBAs may be confounded by the heterogeneity of patients within the clinical environment, with evidence suggesting that case complexity may be a greater predictor of WBA scores than actual participant skill

(Wilkinson et al. 2008). This concept is not unique to medicine, and has been described in the broader field of psychometrics and performance assessment (Sturman et al. 2005). Wilkinson and colleagues found that in addition to case complexity, the testing environment itself may be a significant confounder of test Content validity, with significant score differences seen between the inpatient and outpatient setting (Wilkinson et al. 2008). Threats to the Internal Structure validity of WBAs have been documented in the literature as well. Conflicting findings in the evidence have been found regarding the inter-rater reliability of these assessment types, which may be related to inadequate numbers of assessors, or lack of assessor training (Pelgrim et al. 2011; Kogan et al. 2011). Interestingly, the Internal Structure of these WBAs may be threatened by 'within-person performance variability', manifested as differences in skill or ability of a given trainee or physician within a programme of assessment (Fisher 2015). The scores generated from a WBA may represent that persons 'typical' ability as desired, or rather, it may actually be a reflection of their 'maximum' ability in a given task or procedure. Finally, as with most assessments in surgical education, there remains a global gap in understanding regarding the ability of these assessment scores to predict meaningful clinical outcomes (Kogan & Holmboe 2013).

Faculty buy-in regarding assessment practices has become a huge barrier top the implementation of WBAs in a real-world training program. After a full rollout of a mandatory online assessment tool for surgical trainees in the UK, survey data revealed that the faculty completing these evaluations felt overwhelmingly that WBAs as implemented had very little educational value (Pereira & Dean 2009). Despite these findings, -follow-up survey data revealed that over 5 years, this low opinion of WBAs amongst surgical trainers remained consistent. Of note, assessment data collected during this time found a global misinterpretation of the purpose of this WBA curriculum, with assessors taking a traditional 'summative' approach to these evaluations, typically holding them at the end of rotations and providing little in the way of feedback of critical appraisal (Eardley et al. 2013). These issues are likely the result of trainers feeling these mandatory WBAs were simply 'box-ticking exercises', paperwork that needed to

be completed in order for their trainees to be successful on their career paths, rather than explicit opportunities for learning or remediation (Eardley et al. 2013).

Beyond threats to score validity, there are limitations to WBAs that pertain specifically to a lack of adequate adjustment for clinical or patient features that may make comparison of trainee's performance difficult (Norcini 2005). The most obvious example of this is the simple fact that variations in patient disease presentation make comparisons across multiple assessments or trainees difficult, in addition to potentially requiring a unique set of performance standards for each encounter based on these factors (Norcini 2005). In addition, assessing the patient's experience or outcome is particularly challenging in the context of WBAs, as multiple external factors such as the patient's adherence to treatment plan may in fact be much more significant in determining the outcome-measure, compared to the process being assessed (Norcini 2005). This issue becomes exponentially harder to adjust for when one accounts for differences in rater 'compensation' in the assessment scores, as some judges may be more lenient in their evaluation based on their subjective interpretation of the clinical situation being presented to the trainee (Norcini et al. 2003; Govaerts et al. 2013; Kogan et al. 2011).

Finally, a lack of observational sampling may further confound the use of WBA scores in surgical training. It is imperative that when designing a program of assessment in the workplace, a sampling strategy is used that limits the inherent bias of conducting assessments on large and diverse groups of patients (Norcini et al. 2003). Failure to do so may have implications for the reliability of assessment scores (Van Der Vleuten & Schuwirth 2005), and the generalizability of inferences made from these assessments (Kogan & Holmboe 2013).

10.5 Assessments of Technical Performance

10.5.1 Task-Specific Checklists

The use of checklist-style assessments has been successfully adopted into many high-reliability industries, notably aviation (McCulloch et al. 2009; Lingard 2005). Similar to a Standard Operating Procedure (SOP), standardization through the use of a task-specific checklist (TSC) has potential benefits in reducing the variation in task-execution, and minimizing human error (Lingard 2005). The most notable example of successful implementation of a checklist into surgical practice is the WHO Surgical Safety Checklist (Haynes et al. 2009}, which has been shown to reduce morbidity and mortality of operative procedures in Canada(Urbach et al. 2014) and around the world. The objective, reliable characteristics of a checklist-type assessment has made them popular for use in educational assessment as well, and evidence has supported their use across many specialties and tasks (Ilgen et al. 2015).

Checklists have been used with variable success for the accurate and reliable assessment of trainee and surgeon technical skill. Peyre et al used a Delphi technique to create a Nissen fundoplication-specific checklist for assessing technical skill, and demonstrated excellent internal structure validity of the instrument across five expert raters (Peyré et al. 2010). In a study comparing OSATS GRS score to another Delphi-based, shoulder surgery-specific checklist, Bernard and colleagues found that their checklist had higher interrater reliability than both GRS and overall pass/fail decisions (Bernard et al. 2016). Finally, a methodologically sound study from Harvard Medical School provided multiple sources of validity evidence for a novel central venous catheter (CVC) insertion task-checklist, including setting competency standards using the Angoff method.

TSCs have been used sparingly for high-stakes examinations, such as surgeon credentialing. In urology, the primary example of this comes from Japan, where urologists are assessed on their technical skill in the Endoscopic Surgical Skill Qualificatioon (ESSQ) program (Matsuda et al. 2006). Their system uses error-based

checklist assessments of actual operative footage, submitted by applicants wishing to be certified by this licensing authority. Surgeons certified in this way have been shown to have low level of operative complications, independent of their annual case volume (Habuchi et al. 2011). Their checklist assessment-based approach has shown durable results, as reported in an eight-year follow-up study (Matsuda et al. 2014).

10.5.2 Global Rating Scales

Global rating scales (GRS) have become the predominant manner in which surgical skill is assessed, ranging from formative workplace-based assessments to high-stakes decision making regarding trainee competency (Szasz et al. 2014). These assessment instruments most often employ a Likert-scale approach to rating performance across multiple domains, typically from 1-5 (Szasz et al. 2014). The most notable example of a GRS is the OSATS scale, which assess technical performance across domains such as *Tissue Handling* and *Use of Assistants* (Martin et al. 1997). Often used as a standard for comparison, the OSATS has amassed a huge amount of validity evidence across multiple testing environments (Hatala et al. 2015), including a landmark study demonstrating its ability to discriminate between surgeons with superior and inferior postoperative outcomes (Birkmeyer et al. 2013).

Although some literature would indicate that the reliability of GRS is inferior to that of checklist-based assessments, as mentioned above(Bernard et al. 2016), there have been some studies demonstrating the opposite to be true (Walzak et al. 2015). In a highly touted study from Walzak and colleagues, the internal structure properties of a OSATS-type GRS was contrasted to that of delphi-generated checklists for six bedside procedures. In their analysis, they found that the GRS had higher interrater reliability compared to the procedure-specific checklists, specifically regarding decisions of competence (Walzak et al. 2015). This study lends support the use of GRS-based assessments for making summative decisions around competency in a training program, in particular when used in conjunction with an accepted standard setting method (Szasz et al. 2014).

It is important to note an important distinction relating to the composure of global rating instruments. While most are created to be generic across many surgical procedures and specialties, such as the OSATS(Martin et al. 1997), the Global Objective Assessment of Laparoscopic Skills (GOALS) (Vassiliou et al. 2005), and the Global Evaluative Assessment of Robotic Skills (GEARS) (Goh et al. 2012), newer studies looking at the development of such instruments have taken a more narrow, procedure-specific approach(Kramp et al. 2015; Larsen et al. 2008; Hussein et al. 2016; Raza et al. 2015). This shift has been brought about in a part by a call from surgeons for more nuanced, focused feedback, specifically regarding the manner in which the operative step was executed (Ghani et al. 2016). Additionally, there is some evidence to suggest that higher reliability is achieved when using an assessment scale with procedure-specific domains and anchor points (Ghani et al. 2016; Kramp et al. 2015).

10.5.3 Safety Metrics

While not classically considered measures of skill or ability, recent evidence supports the use of technical errors and intraoperative adverse events (iAEs) as methods of assessing trainee performance (Bonrath, Dedy, Zevin & Grantcharov 2013a; Rogers et al. 2002; Sarker et al. 2005). An early example of this approach comes from the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) program, where performance scores are a composite measure that includes time to task-completion and the number of errors committed by the trainee (termed ‘precision’) (Fried et al. 2004). This type of assessment has been adopted by the Fundamentals of Laparoscopic Surgery (FLS) program, which is used for both credentialing and continuing medical education (CME) purposes in many parts of the United States (Swanstrom et al. 2006). Evidence has shown that implementing error-based assessments into training have positive effects on trainee technical skill (Sroka et al. 2010; Fried et al. 2004).

Methods of quantifying and classifying technical errors and iAEs have been developed for assessing technical skill in the operating room as well. Interestingly, these assessments have been done in large part in laparoscopic cholecystectomy cases, over the past 20 years. Joice and colleagues initially used a task-analysis strategy to assess surgeon technical errors in laparoscopic cholecystectomy, based on operative video and written description of the procedure (Joice et al. 1998). They found a huge number of iatrogenic injuries to the gallbladder were the result of avoidable technical errors, among the 20 cases they observed. Although they did not formally 'assess' the surgeons in an educational sense, or provide structured feedback using their task analysis-style evaluation, this study revealed the breadth and scope of surgeon technical errors in even seemingly straightforward operations (Joice et al. 1998). Tang and colleagues carried this work forward in the simulation environment to assess trainee performance, showing that HRA can be used to categorize and classify errors committed by surgical residents, to better direct training goals and curricular design (Tang et al. 2006; Tang et al. 2005). In a study that quickly followed Joice's, Eubanks et al used a similar method of task-analysis, based on expert consensus, to create a procedure-specific list of common errors committed during the operative steps of a laparoscopic cholecystectomy, weighting each error by their perceived impact on patient safety (Eubanks et al. 1999). They provided good internal structure validation of their novel scoring system, however it failed to adequately capture the variation across surgeons of different training levels. Finally, a London-based group led by Lord Darzi categorized errors and adverse events during laparoscopic cholecystectomy, again using expert consensus (Sarker et al. 2005). They found that expert surgeons make significantly fewer major technical errors, compared to minor ones, and propose a theoretical model regarding fluctuations in error occurrences based on their data (Sarker et al. 2005).

In the last five years, our understand of technical errors and adverse events in surgical assessment has changed significantly, in large part through work from Bonrath et al (Bonrath, Dedy, Zevin & Grantcharov 2013a). They identified a global need for a standardized set of definitions around surgical error and iAEs, to drive research and

develop novel methods of quantifying risks to intraoperative patient harm (Bonrath, Dedy, Zevin & Grantcharov 2013a; Bonrath, Dedy, Zevin & Grantcharov 2013b). Their work went beyond the identification and classification of surgical errors and events (Bonrath & Gordon 2015), by developing and providing key validity evidence for a novel assessment tool that demonstrated discriminatory ability among trainees and faculty surgeons with high and low OSATS scores during laparoscopic Roux-en-y bypass procedures (Bonrath, Zevin, et al. 2013). This tool was further analyzed in the context of laparoscopic hysterectomy, with these findings replicated (Husslein et al. 2015).

10.5.4 Video-Based Assessments

10.5.4.1 Rationale

As highlighted in this thesis, surgery remains far behind other high-reliability professions regarding the use of technology-enhanced education and credentialing practices. Athletics in particular has spearheaded much of the work in video-based review, investing huge amounts of resource and time into the development of novel ways of using this rich data source to enhance individual and team performance (Sarmento et al. 2014). Watching the 'game-tape' has become ubiquitous across nearly every professional sport, with organizations hiring experts in video analysis as part of their management or coaching groups (Haynes et al. 2009)
(Toner & Moran 2014). A similar phenomenon has been seen in teacher training, where a wealth of literature has been devoted to understanding the benefits of using a video-based strategy to enhance the competency of child educators (Christ et al. 2017; Wiens et al. 2013; Admiraal et al. 2011). However, despite the apparent benefits of incorporating video review into surgical training and practice (Aggarwal et al. 2008), there has not yet been a similar investment into this technology as seen in other industries.

Understanding variations in healthcare quality is a complex task, made more difficult by the often inadequate granularity provided by institutional or population-level data (Dimick & Greenberg 2013). Video-documentation of medical or surgical

procedures provides a rare opportunity to better understand the variations in technique that exist between providers, and help to standardize the way in which these tasks are carried out (Rex et al. 2010). Furthermore, keeping a prospective, objective record of a patient's journey through the healthcare system can enhance our ability to study and learn from system and physician errors that create unsafe care and lead to adverse events (Yang et al. 2016). The recall bias associated with Morbidity and Mortality Conferences (MMC) can be minimized by augmenting case review with video data from the patient's procedure(Kjellsson et al. 2014). Video can further minimize threats to patient safety by shortening the learning curve associated with complex procedures (Ibrahim et al. 2016). Finally, video has the potential to improve the always fragile relationship between patients and their physician or hospital (Makary 2013). Two US state legislatures have previously debated passing into law a requirement for hospitals to provide patients with a video recording of their surgical procedure (Legislature 2015), spurred by stories recently covered in the media relating to mistreatment of patients in the operating room (Leverage 2015).

10.5.4.2 Applications in Education

Traditionally, evaluations of surgical performance and safety have relied on live assessments by human observers (Williams et al. 2016). While this has been shown to provide accurate and reliable assessments in training (Kogan et al. 2009), there are inherent limitations when a rater is present at the time of the assessment. Most notably, the inability to blind the assessor to the identity or experience level of the trainee can introduce unwanted bias into the assessment process (Williams et al. 2016; Dagnaes-Hansen et al. 2017). Additionally, video-based review allows for the evaluation process to be re-watched as needed, thereby limiting the inaccuracy associated with performance assessments completed at a later time (Williams et al. 2014).

Video-recorded assessments in surgical education have been shown to be feasible and practical (Aggarwal et al. 2008). In a seminal article by Beard and colleagues(Beard 2005), the saphenofemoral disconnection step of a vascular surgery

procedure was video-captured and assessed using the OSATS scale, with the ability of surgeons across different levels of skill and experience analyzed. Twenty-eight surgeons (14 trainee, 14 staff-level) were included, and participants assessed each other's performance blindly using the video recordings. This volume of assessment is likely impossible without the ability to capture and analyze video data in the operating room. De Montbrun et al studied the technical skills of first year general surgery residents over ten years, and set procedural standards using this cohort (de Montbrun et al. 2015). This high volume of assessments over such a long study period, improves the generalizability of their findings, and ensures that you can confidently make decisions regarding the competency of a given trainee. By recording technical performance, whether in the operating room or in the laboratory, educators are able to ensure they not only meet the number of assessments required to answer their research questions, but also that they are able to ensure that enough judgments are collected to be confident in the establishment of assessment standards.

While omnipresent in athletics, the concept of coaching has only recently found its way into the repertoire of surgical educators. The use of video-recordings of surgical procedures has allowed for postoperative coaching sessions to occur (Greenberg et al. 2016). While this would typically an experienced surgeon providing a trainee with feedback specific to their performance during a procedure (Soucisse et al. 2016), peer-to-peer coaching has be used to allow two established surgeons to share and learn from one another (Greenberg et al. 2017). Video review in the coaching process allows for a surgeon to watch their own performance with a coach and receive feedback on various factors relating to their execution of a procedure (Yanes et al. 2016). This enhances the feedback process, as those being assessed are able to better understand how evaluators reached certain conclusions regarding their performance (Yanes et al. 2016). This process has recently been identified as a possible means of improving surgical quality in urology specifically, with coaching being used as a method of increasing operating room teaching, and improving intraoperative judgement and technique (Penson 2012).

10.5.4.3 Limitations and Barriers to Implementation

When discussing the use of video for the evaluation of surgical technique and skill, it is important to remember the limitations, both ethical and practical, associated with this type of data collection. Data from the operating room must be collected in such a way as to preserve patient autonomy, safety, and privacy (Langerman & Grantcharov 2017). Langerman categorize the types of data collected in the operating room into procedural video, panoramic video and audio, and digital data(Langerman & Grantcharov 2017). Each of these sources of patient and surgical team data must be protected in the same manner as all quality improvement data, with participant informed consent, deidentification where feasible and possible, and alignment with institutional and locoregional laws and practices. Furthermore, they highlight the potential ethical implications for the surgeon and surgical team, in particular the role of the Hawthorne Effect, conflicts between recordings and the patient record (i.e. the operative note), and the issues around a surgeon's handling of missed intraoperative errors identified through retrospective review of the video. Additionally, one must consider the medicolegal implications of recording data in the operating room for education and quality initiatives, and jurisprudence internationally speaks to the importance of ethical practices when recording, storing, and utilizing these data (Henken et al. 2012). Strategies to decrease the legal risk of organizations undertaking these initiatives includes storing data for set periods of time only and erasing previously analyzed data, and classifying operating room data as in such a way as to make it non-discoverable in legal proceedings, in the same way that other institutional medical records and outcome data are stored (Boer et al. 2018; Henken et al. 2012). Defining a clear use for this type of data may reduce the likelihood that it be used for punitive reasons rather than quality improvement as intended (Prigoff et al. 2016). While collecting data in the operating room provides unique ethical challenges for educators and researchers, the potential to enhance safety and improve efficiency make data capture in the operating room a worthwhile endeavour.

10.5.5 Assessments in Robotic Surgery

10.5.5.1 Implementing Assessments of Robotic-Assisted Technical Skill in Urologic Education: A Systematic Review and Synthesis of the Validity Evidence

10.5.5.1.1 Introduction

Surgical education is experiencing a huge shift from Halstead's apprenticeship model introduced over 100 years ago to the current climate of competency-based education. A trainee must exhibit clinical competence, and in surgical education this includes both the technical and non-technical skills needed to safely carry out any number of procedures. Evidence linking technical performance to patient outcomes and safety has drawn the public's attention, reflected by recent efforts to allow patient access to video footage of surgical procedures(Langerman & Grantcharov 2017). These developments have significantly altered the way we in which approach research in surgical assessment and curriculum design.

More than in any other surgical field, robotic-assisted surgery (RAS) has been rapidly embraced by the urologic community. It is quickly becoming the most common approach to many operations, including prostatectomy, partial nephrectomy, pyeloplasty, cystectomy, and retroperitoneal node dissection (RPLND)(Zorn et al. 2009). Its predominant use continues to be for prostate cancer, where Robotic-Assisted Radical Prostatectomy (RARP) has become the gold standard in the surgical management of localized prostate cancer in most of the developed world(Zorn et al. 2009). The dynamic growth of this surgical technique has had a wide impact on practicing urologists and surgical residency and fellowship programs alike. The need for formalized RAS training has also resulted in increased need for assessments of skill, both formative and summative. Despite the continued creation of new tools to assess performance, important questions remain unanswered; how do we effectively incorporate RAS training programs into urology residency curricula? How do we appropriately credential practicing urologists wishing to perform robotic surgery?

How do we incorporate the most effective education programs in the urology residency curricula? Most recent Urology residency graduates will not have had an immersive experience in robotic surgery. Those Urologists who passed their Board or Fellowship exams 10 years ago have had to acquire the required robotic skills in a very unstructured transitioning surgical landscape. It may even be appropriate to include robotic surgical education curricula late in medical school training. This would permit early recognition of those students with aptitude in surgery to be identified using the metrics outlined in this manuscript.

For an objective assessment of robotic skill to be applicable in training, privileging or accreditation, it is essential to build a 'validity argument' supporting its use. Messick's Conceptual Framework is an acceptable way to construct such an argument, through the assembly of various sources of validity evidence, specifically *content, response process, internal consistency, relationship to other variables,* and *consequences*(Messick 1994). This type of framework replaces the now outdated Cronbach Taxonomy of validity (predictive, concurrent, content and construct validity), by seeing validity as a dynamic or fluid concept that must be argued in different assessment environments.

Like any procedural assessment rubric, the tools used to evaluate robotic skill employ a combination of global rating scales (GRS) and task-specific checklists to assess trainee competencies(Goh et al. 2012; Siddiqui et al. 2014). Using trained expert analysts, GRS can be superior in both accuracy and reliability across a wide variety of procedure-types when compared to checklists(Regehr et al. 1998). Despite this the validity evidence supporting objective assessments of technical skill remains insufficient to warrant their use in high-stakes decisions such as progression through competency-based training or credentialing(Hatala et al. 2015). It is vital to create a validity argument in support of these approaches when considering their inclusion in summative assessments in training and beyond(Kane et al. 1999).

While both technical and non-technical skills are essential in the training of future robotic surgeons(Tiferes et al. 2016),this article focuses on technical skill assessments only. The objective of this article is to provide a focused review of the available tools for assessment of robotic surgical technical skill currently available to surgeon educators, and to critically appraise the supporting literature to determine how best to implement these assessment tools into residency and fellowship curricula.

10.5.5.1.2 Methods

Eligibility criteria

Articles assessing the robotic surgical skill of urologic trainees (medical students, residents, fellows) or faculty urologists were included. Studies assessing robotic skills in other surgical specialties, that did not include urology participants were excluded. Studies primarily assessing non-technical skills were excluded from this review, although the search was designed to capture these studies for future work. Studies published in peer-reviewed journals were included in the analysis, and unpublished abstracts were included only if it was determined that they contained data contributing to the validity of the assessment being studied. Randomized control trials and observational studies, including cohort, case–control, case series and cross-sectional studies, were all eligible for inclusion.

Information sources

One author conducted a search in Ovid MEDLINE, Embase Classic, PsycINFO and the Cochrane Library. The search was carried out on July 18th, 2017.

Search

Medical subject headings (MeSH) terms used in the search included ‘communication’, ‘clinical competence’, ‘curriculum’, ‘education, medical’, ‘surgical

procedures', 'education, medical, graduate', 'educational measurement', 'medical errors', 'nephrectomy', 'patient simulation', 'prostatectomy', 'robotic', 'robotic surgery', 'robotic surgical procedures', 'robotics', 'skill', 'surgery', 'non-technical skill', 'cognitive skill', 'technical', 'technical skill', 'urologists', 'urology'. Titles of articles resulting from the search and corresponding abstracts were reviewed initially and articles eligible for full-text review were identified. These articles were then analyzed further to ensure that no articles referenced therein were missed for inclusion in the full-text review. Duplicates were identified and removed.

Study selection

Any study in the medical or surgical literature that assessed the robotic surgical skill of urologic trainees or faculty, involving original research and described in English, were included. Opinion letters, editorials, case reports, reviews, and letters to the editor were excluded. References used in previous review articles were assessed and those that met the inclusion criteria were incorporated in the analysis. Articles that looked at outcomes only were also excluded. Two authors considered the articles for inclusion independently, and any disagreements were resolved by consensus.

Data collection process

Data were abstracted from the included studies systematically, including sample size, participants, assessment used, study setting, rater information, and assessment design and implementation relevant to various sources of validity evidence.

Quality assessment

The Medical Education Research Study Quality Instrument (MERSQI) was used to assess the quality of the included articles (Cook & Reed 2015). The MERSQI scores quality over eight domains: study design, institutions sampled, response rate, type of

data, validity evidence for evaluation of instrument scores, sophistication of data analysis, appropriateness of data analysis, and assessment outcome.

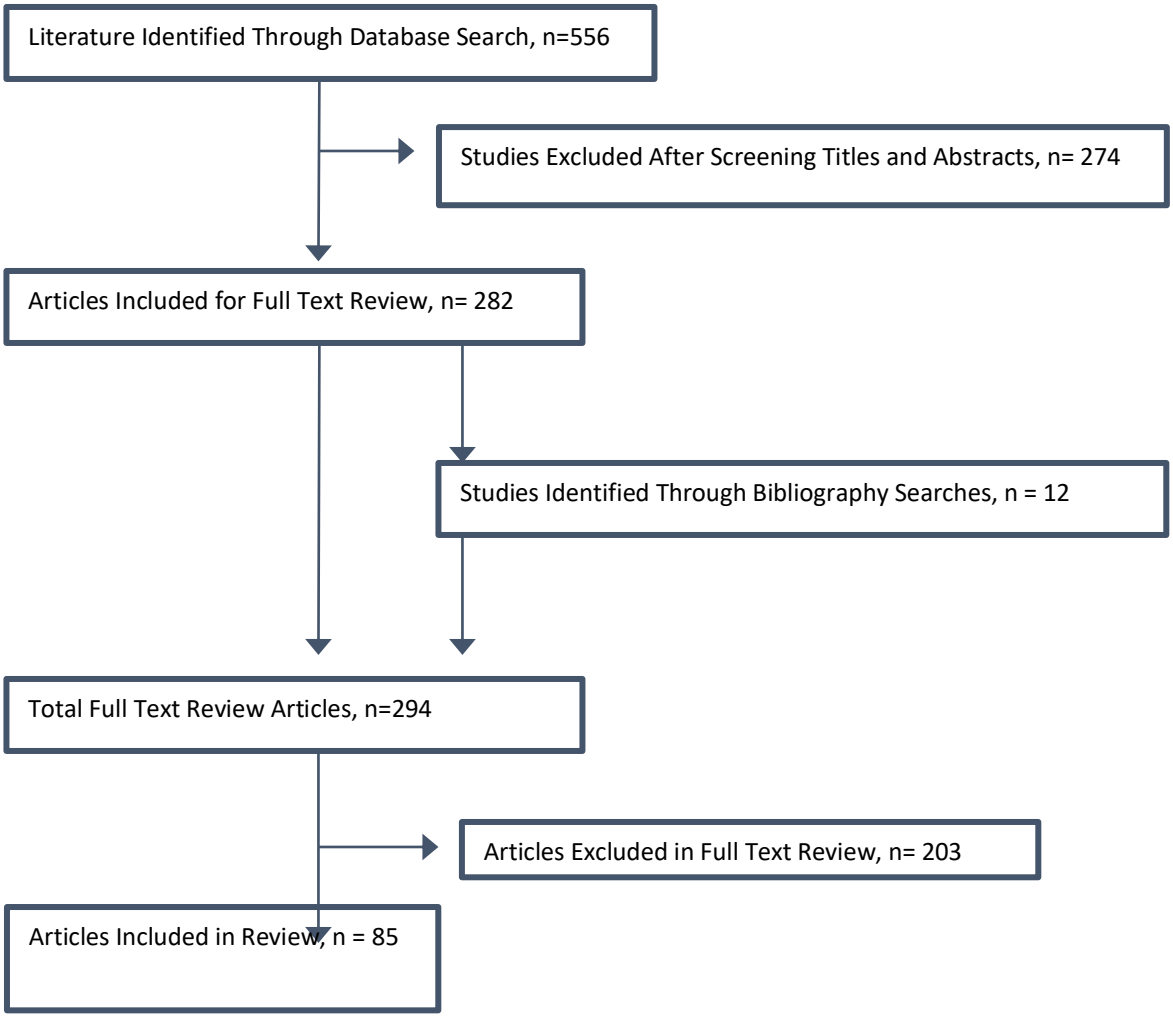
Validity Evidence

We used Messick's validity framework(Messick 1994) to structure the evidence supporting the application of these assessment tools in robotic surgery. These sources of test validity include content, response process, internal structure, relationships to other variables, and consequences of testing. Use of this framework allowed us to put forward our own, evidence-guided recommendations on how best to implement these assessments into formal training curricula.

10.5.5.1.3 Results

Our initial search yielded 566 articles. After two independent authors reviewed titles and abstracts, 282 articles were selected for full review to determine inclusion status. Following full text review and cross-checking of article references, 85 studies were included in the final analysis (Figure-1). The included articles are displayed in Appendix-1, subdivided into assessments of technical skill and computer-based virtual reality (VR) assessments.

Figure 1 PRISMA Flow Chart



Technical Skill Assessments

Table-1 summarizes the validity evidence supporting the seven non-time-based technical skill assessment tools used in urological robotic surgery. The Global Evaluative Assessment of Robotic Skills (GEARS) tool, developed by Goh et al, has been applied to urological assessments on multiple occasions(Vernez et al. 2016; Mills et al. 2017; M. G. Goldenberg, L. Goldenberg, et al. 2017; Hung et al. 2013; Goh et al. 2012; Ramos et al. 2014; Ghani et al. 2016; Powers et al. 2015; Gomez et al. 2015; Aghazadeh et al. 2015; Aghazadeh et al. 2016; Hung et al. 2017; Volpe et al. 2015; Hung et al. 2015; Holst, Kowalewski, White, Brand, Harper, Sorenson, et al. 2015; White et al. 2015; Chowriappa et al. 2015; Whitehurst et al. 2015; Holst, Kowalewski, White, Brand, Harper, Sorensen, et al. 2015; Dubin et al. 2017), and has the strongest validity argument supporting its use in the assessment of robotic skill. Its generic framework has allowed it to become a widely accepted method of assessment across multiple procedures and even across specialties(Aghazadeh et al. 2015; Ramos et al. 2014; Goh et al. 2012). Notably, evidence supports its ability to discriminate amongst staff surgeons of differing case volume(Ghani et al. 2016), as well as across a single surgeons learning curve(M. G. Goldenberg, L. Goldenberg, et al. 2017). The vast majority of literature using the GEARS score has found it to be a reliable assessment method(White et al. 2015; Holst, Kowalewski, White, Brand, Harper, Sorensen, et al. 2015; Goh et al. 2012; Ghani et al. 2016; Vernez et al. 2017; Whitehurst et al. 2015; Holst, Kowalewski, White, Brand, Harper, Sorenson, et al. 2015). However, a study of robotic renal hilar dissection using oriented expert raters showing poor internal consistency(Powers et al. 2015), and Hung and colleagues found that while trainee self-assessments and faculty evaluations correlated weakly, inter-faculty reliability was better when assessing residents (ICC=0.77) and fellows (ICC=0.45)(Hung et al. 2017). As shown in Table-1.1, it is the only technical skill assessment tool that has supporting *consequences* evidence, having been used to both predict clinical outcomes in a retrospective case-control study, and impact residency match-rankings when applied to a cohort of medical students.

Table-1 Validity Evidence for Assessments of Technical Skill

						Messick's Framework of Validity				
	Instrument Description	**Domains Assessed**	**Number of Studies as Primary Assessment Method**	**Number of Participants, Primary Assessment Method**	**Number of Studies as Secondary Assessment**	**Content**	**Response Process**	**Internal Structure**	**Relationship to Other Variables**	**Consequences of Testing**
Technical Skill Assessments										
GEARS	Robotic-specific GRS; expansion of GOALS with expert consensus	Depth perception Bimanual dexterity Efficiency Force Sensitivity Autonomy Robotic control	18	569	2	17	11	11 IRR range 0.38-0.92 (M=0.80)	12	2 1. Scores used to determine ranking 2. GEARS score predicts surgical outcome
OSATS	GRS; developed initially for assessing basic surgical skills in OSCE examination	Respect for Tissue Time and Motion Instrument Handling Knowledge of Instruments Flow of Procedure Use of Assistants Knowledge of Procedure	7	345	1	9	4	3 IRR range from 0.84-0.91 (M=0.87)	8	0

GOALS	GRS; developed to assess laparoscopic skill	Depth perception Bimanual dexterity Efficiency Tissue Handling Autonomy	3	72	1	4	3	2 IRR 0.66-0.80	3	0
R-OSATS	GRS; four dry-lab exercise-specific scale that combines elements of GOALS and OSATS	Depth Perception Force Sensitivity Dexterity Efficiency	1	105	0	1	1	1 IRR 0.79	1	0
PACE	Procedure-Specific GRS for RARP	Anchored Likert Scale Across 7 Operative Steps	1	56	0	1	1	1 IRR ranged from 0.4-0.8 across domains	1	0
ARCS	Robotic-specific GRS developed by Intuitive Surgical technician-trainers	Dexterity Optimizing Field of View Instrument Visualization Optimizing Workspace Force Sensitivity and Control Basic Energy Pedal Skills	1	15	0	1	1	1 IRR ranged from 0.52-0.81 across domains	1	0

RACE	Task-Specific GRS developed to evaluate urethrovesical anastomosis performance	Needle Positioning Needle Entry Needle Driving & Tissue Trauma Suture Placement Tissue Approximation Knot Tying	2	40	0	2	1	1 IRR 0.55-0.62	1	0
RARP Assessment Score	Prostatectomy-Specific Assessment based on HFMEA analysis	Operative steps broken down into sub-steps with hazard categories assigned for modular introduction to RARP	1	15	0	1	1	1 Kappa range - 0.241 – 0.2 Significant agreement on 2/27 steps	1	0

The Global Operative Assessment of Laparoscopic Skills (GOALS) (Xu et al. 2016; Hung et al. 2012; Tunitsky et al. 2013; Vernez et al. 2016), a laparoscopic-specific GRS that served as the underlying model for the GEARS, was also used in robotic skills assessment by Hung et al(Hung et al. 2012), with the addition of two robotic-specific domains, *instrument awareness and precision and camera awareness and precision*. Their randomized control trial demonstrated that baseline performance on a virtual reality simulation scenario correlated with performance on a porcine model. Tunitsky et al(Tunitsky et al. 2013) demonstrated GOALS ability to discriminate between *procedural expert* surgeons and *robotic expert* surgeons performing a simulated robotic ureteral anastomosis, providing evidence that this GRS may be able to adequately evaluate procedural-specific constructs. The Objective Structured Assessment of Technical Skill (OSATS) tool(Vlaovic et al. 2008; Korets et al. 2011; Tarr et al. 2014; Gomez et al. 2015; Vernez et al. 2016; Phé et al. 2017; Alemozaffar et al. 2014; Rashid et al. 2006), originally developed at the University of Toronto for a 'bench-station' examination of basic surgical skills(Reznick et al. 1997), has been used to assess robotic technical skill, with multiple studies providing various types of validity evidence, across simulation, laboratory, and clinical environments. Siddiqui and colleagues(Siddiqui et al. 2014) added robotic-specific metrics to the OSATS tool, using 5 dry-lab 'drills' to assess robotic skill across 4 domains, terming their modification 'R-OSATS'. They demonstrated its relationship to other variables by comparing scores to training level and console experience. Their tool also exhibited excellent inter-rater reliability (Cronbach's $\alpha = 0.91$). A RARP-specific assessment tool, the Robotic Anastomosis Competency Evaluation (RACE)(Raza et al. 2015; Ghani et al. 2016) was developed by Raza et al, and uses global ratings across 5 domains to assess specific skills needed to complete the vesicourethral anastomosis step of the RARP. While their tool could discriminate between trainees of different experience, the reliability of their tool was only moderate ($\alpha = 0.62$). The RARP Assessment Score(Lovegrove et al. 2015) was developed by an international group using the Healthcare Failure Mode Effect Analysis (HFMEA). The HFMEA(L LA PIETRA 2006) is a method of human risk analysis, which allowed the authors to identify high-risk steps of the procedure to include in their assessment of trainees taking part in a European robotics fellowship. However, the

small numbers of participants in their study makes interpretation of their data difficult at this stage. The Prostatectomy Assessment Competency Evaluation (PACE) is the product of a Delphi consensus of international urologic oncologists(Hussein et al. 2016). Like the RARP Assessment Score, this tool is procedure specific. Each step of the procedure is rated using a 5-point Likert scale, with agreed upon anchor points for scores of 1, 3 and 5. Finally, the Assessment of Robotic Console Skills (ARCS) was developed in collaboration with Intuitive Surgical as a global rating scale to more-specifically assess console skills, including optimization of field of view and workspace, and basic energy pedal skills(Liu et al. 2017). Their initial validation study demonstrated the ARCS ability to discriminate between staff surgeons of less than 100 versus greater than 100 completed robotic-assisted cases.

In addition to these GRS assessments, studies used weighted combinations of time and error(McVey et al. 2016; Goh et al. 2015; J. Y. Kim et al. 2015; Hinata et al. 2013; Arain et al. 2012; Lendvay et al. 2013; J. Y. Lee, Mucksavage, Kerbl, et al. 2012; Tausch et al. 2012; Dulan et al. 2012; Stegemann et al. 2013; Davis et al. 2010; Amirian et al. 2014; Foell, Finelli, et al. 2013) (similar to the Fundamentals of Laparoscopic Surgery, FLS(Fried et al. 2004)) and 'end-product' scores[50,58,59,62] to assess technical performance.

Computer-Based VR Assessment

Table-2 outlines the commercially available simulation platforms and scoring metrics for robotic surgery with literature supporting their use in training urologists. The field of robotic simulation is well established, with multiple developers offering platforms to the public, each with its own unique features, strengths and weaknesses(Moglia et al. 2015).

Table-2 Validity Evidence for Computer-Based Virtual Reality Assessments

						Messick's Framework of Validity				
	Instrument Description	**Domains Assessed**	**Number of Studies as Primary Assessment Method**	**Number of Participants, Primary Assessment Method**	**Number of Studies as Secondary Assessment**	**Content**	**Response Process**	**Internal Structure**	**Relationship to Other Variables**	**Consequences of Testing**
Simulation-Based Assessments										
dV-Trainer/ MdVT	Computer-Generated metrics developed by Mimic Simulation	Time Economy of Motion Drops Instrument Collisions Excessive Instrument Force Instruments Out of View Master Workspace Range	12	525	1	12	2	0	8	0
dVSS	Computer-Generated metrics developed by Intuitive Surgical	Camera targeting Energy switching Threading rings Dots and Needles Ring and rail	23	697	3	26	12	0	21	2 1. dVSS scores predict GEARS score in OR 2. dVSS scores predict performance on dry-lab tasks using robotic console

RoSS (RSA-Score)	Computer-Generated metrics developed by the University of Buffalo and the Roswell Cancer Institute	Task Time Safety in Operative Field Economy Bimanual Dexterity Critical Errors	2	57	0	2	0	1 Internal Domain Consistency 0.01-0.98	1	0
SEP	Simulator developed in the Netherlands		2	63	0	2	0	1 IRR 0.73	1	0
RobotiX	Computer-Generated metrics developed by Simbionix Products	Fundamentals of Robotic Surgery and Robotic Suturing Modules	1	46	0	1	0	0	1	0
ProMIS	Adapted from Laparoscopic Training System from Haptica (Ireland)	Peg Transfer Precision Cut Intracorporeal Knot	3	73	0	3	0	0	3	0

Intuitive Surgical (Sunnyvale, CA), designer of the daVinci System, is responsible for the daVinci Surgical Simulator (dVSS)(Aghazadeh et al. 2016; Wiener et al. 2015; Volpe et al. 2015; G. I. Lee & M. R. Lee 2017; Noureldin et al. 2016; Hung et al. 2013; Meier et al. 2016; D. C. Kelly et al. 2012; Finnegan et al. 2012; Dubin et al. 2017; Hung et al. 2011; Foell, Furse, et al. 2013; Brown et al. 2017; Yang, Zhen, et al. 2017; Lyons et al. 2013; Mark et al. 2014; Yamany et al. 2015; Hung et al. 2012; Liss et al. 2015; Alzahrani et al. 2013; Hassan et al. 2015; Liss et al. 2012; Ramos et al. 2014; Korets et al. 2011; Song & Ko 2016; Lendvay et al. 2013). This robotic simulator fits directly onto the surgeon console, allowing the trainee to sit at the same controls he or she would be using in the operating room. It has the disadvantage of not being available if the console is being used in the operating room, as it cannot be used independently of the console(Tanaka et al. 2015). The dVSS is the result of collaboration. The software used by the dVSS was developed initially in conjunction with the Mimic group, and so many similarities are found between these platforms in terms of metrics assessed and the user interface (UI). In 2009, Lerner and colleagues(Lerner et al. 2010) showed that a cohort trained on the dV-Trainer® performed similarly to those trained on the dVSS, and they achieved similar results on dry-lab tasks. This outcome may reflect the similarities in their software design and UI. Additionally, the selection and creation of the tasks used by the dVSS was made in conjunction with the Simbionix group. In a study by Amirian et al(Amirian et al. 2014), the Simbionix suturing module (SSM), running on the dVSS training software, was able to demonstrate improvement from baseline in a group of robotic novices. Lee et al developed a four-week training curriculum, the Basic Skills Training Curriculum (BSTC)(J. Y. Lee, Mucksavage, Canales, et al. 2012), which employed the dVSS system to compare a time-based method of assessment with a competency/proficiency-based method in surgeons of various training levels at the University of Toronto. Hung and colleagues(Hung et al 2011) used visual analogue scales (VAS) to establish the functional task alignment of the dVSS, and their study showed again that this simulation platform can distinguish between experts and novices. In a subsequent study, this group demonstrated that assessments with the dVSS have clinical consequences(Hung et al. 2012), by correlating baseline trainee skill with ex vivo tissue performance after the completion of a dVSS dry-lab curriculum.

Another popular robot-specific platform is the dV-Trainer® developed by Mimic (Seattle, WA) (Yang, Zhen, et al. 2017; Raison, Ahmed, et al. 2017; Kang et al. 2014; Perrenot et al. 2012; Kenney et al. 2009; Lendvay, Casale, Sweet & C. Peters 2008a; Sethi et al. 2009; Lerner et al. 2010; Schommer et al. 2017; Ruparel et al. 2014; FACS et al. 2013; J. Y. Lee, Mucksavage, Kerbl, et al. 2012). Initial validation studies(Lendvay, Casale, Sweet & C. Peters 2008b; Kenney et al. 2009) provided evidence that the simulator was able discriminate between expert and novice robotic surgeons. In a 2012 study, Lee and colleagues(J. Y. Lee, Mucksavage, Kerbl, et al. 2012) demonstrated that dV-Trainer® performance correlates with actual daVinci console performance at dry-lab tasks. New initiatives from Mimic include the Xperience Team Trainer, which includes an assistant laparoscopic simulator that integrates a communication element into the simulation experience.(Xu et al. 2016)

Simbionix (Israel) has developed multiple procedural simulators across different specialties, including the RobotiX Mentor Platform®. Like the dV-Trainer®, it too is a stand-alone platform and can incorporate a laparoscopic assistant simulator. Validity evidence for its use comes from a study from Whittaker and colleagues(Whittaker et al. 2016), in which they were able to demonstrate significant score differences between novices and experts, using two simulated modules and employing domains of assessment from the Foundations of Robotic Surgery curriculum (FRS). Simbionix-developed software that allows trainees to complete virtual reality steps of the radical prostatectomy have been recently integrated into both the RobotiX and dVSS platforms.

The Robotic Surgery Simulator (RoSS) (Chowriappa et al. 2013; Seixas-Mikelus et al. 2010), made by Simulated Surgical Systems (San Jose, CA), is another simulator, and unlike the dVSS, it is a standalone platform. While it is not identical to the daVinci console used by the dVSS, it is modeled after it, and subsequently has similar task alignment(Seixas-Mikelus et al. 2010). It was developed with the Roswell Park group in Buffalo, NY, and this group has demonstrated that the RoSS has the ability to predict performance on another simulator(Kesavadas et al. 2009), as well as intraoperative ability(Guru et al. 2009). Finally, the RoSS simulator has now integrated the RSA-score

assessment tool(Chowriappa et al. 2013), developed through the FSRS group as described above, further adding to its applicability to robotic curricula.

The final platform designed specifically for robotic surgery simulation is the Sim surgery Education Platform (SEP) Robot Simulator (Oslo, Norway). This is a less utilized platform, and the evidence for it has been mixed(Gavazzi et al. 2011; Shamim Khan et al. 2013; Balasundaram et al. 2008). Studies have been able to show that novices performed consistently poorer when compared with a cohort of experts on the SEP platform.

A unique example of laparoscopic simulator technology being applied to robotic surgery is the ProMIS system(McDonough et al. 2011; Jonsson et al. 2011; Chandra et al. 2010): a platform that measures efficiency of task completion such as total distance of instrument arm movements and smoothness of motion(McDonough et al. 2011). A urology-specific example of its use in robotics comes from a study by Jonsson et al(Jonsson et al. 2011), who's group showed that the ProMIS simulator was able to discriminate between novices and experts at a dry-lab vesicourethral anastomosis model. This article further added to its validity evidence by comparing the smoothness of motion metric between groups, to the more conventional measurement of time to task completion.

Key differences exist between these simulators. A unique and important property of computer-based VR simulators is the ability to automatically track instrument movements. The dV-Trainer® and SEP simulators measure the force with which the instruments are used, as well as instrument collisions, an important issue with robotic surgery where haptic feedback does not exist. The dVSS contains the 'system settings' and 'wrist manipulation' measurements, performance domains specific to RAS. Interesting assessments incorporated into the SEP platform are tightening and winding stretch. These measure the amount of tension used in knot tying, an important and advanced robotic skill. Finally, the Mscore assessment rubric developed by Mimic and incorporated into the dV-Trainer (older versions of Mscore also found on the dVSS)

allow surgeon mentors and educators the ability to individualize training curricula with development of customized tasks and modular learning activities and deliberate practice sessions based on trainee needs.

Novel Assessment Methodologies

Novel methods of assessing robotic surgical skill have been introduced in the recent literature. We describe four such innovations here, and they are summarized in Table-3.

Table-3 Novel Methods of Assessing Robotic Skill

Assessment Method	**Description of Innovation**	**Levels of Training**	**Setting of Assessment**	**Advantages of Method**
Crowdsourced Assessments	Enlists large numbers of people via an internet platform to complete assessments of technical skill	Medical Students Residents Fellows Staff	Dry-Lab Simulation Wet-Lab Operating Room	Rapid, high volume assessments of video High interrater reliability statistics
Machine Learning	Automated analysis of master workspace adjustment, camera manipulation, unsafe motion and collisions	Residents Fellows	Dry-Lab	Automated analysis of surgeon psychometrics Excellent classification accuracy Potential for real-time, high reliability assessment of performance
Contact Vibrations	Use of contact vibrations, applied force, and time to completion as measures of clinical skill	Staff	Dry-Lab	Improvement classification accuracy of a global rating scale assessment of technical skill

Assessment Method	Description of Innovation	Levels of Training	Setting of Assessment	Advantages of Method
Armrest Load	Use of a pressure surveillance system to detect armrest load on the robotic console	Medical Students	Simulation	Use of pressure-alarm in training can improve ergonomic positioning in novice surgeons Potential for shortening of learning curve in novice trainees

Crowdsourcing

An exciting but controversial area of assessment being established in robotic technical skill assessment is ‘crowdsourcing’(White et al. 2015). This method uses members of the public, medically trained or not, to make judgments on surgical skill and technique. Consistently, studies have shown that these groups of people, often referred to as ‘turkers’, have not only excellent internal consistency, but also have ratings correlative to those of expert surgeons(White et al. 2015). C-SATS(Holst, Kowalewski, White, Brand, Harper, Sorensen, et al. 2015), an online platform that utilizes this method, has been used in multiple surgical fields, including laparoscopy and robotics. Recently, efforts from the Michigan Urological Surgery Improvement Collaborative (MUSIC) have applied this method of assessment to robotic radical prostatectomy(Ghani et al. 2016), showing that crowdsourcing is applicable to assessment of this procedure using GEARS. However, it was noted that the ‘crowd’ was less willing to rate participants as either very poor or very good performers, which was not the case for expert raters. This phenomenon may question the use of this method in summative or high-stakes assessments, where distinguishing between high and low performers is imperative. Additionally, there is a considerable cultural barrier to overcome in this case, as experienced surgeons may doubt the ability of non-medically trained crowd workers to potentially judge whether surgeons are competent at performing advanced surgical procedures. Certainly, there will be more investigation into this assessment modality, including whether crowd-derived judgments can reliably predict not only expert opinion but also patient outcomes.

Machine Learning

A study by Kumar et al(Kumar et al. 2012) used a form of artificial intelligence (AI), Support Vector Machines (SVM), to assess the robotic workspace adjustment and camera manipulation of trainees performing a variety of tasks on the robotic console. They found that their algorithm had a classification accuracy of over 95% for workspace

adjustment, and over 88% for camera manipulation. Despite some study limitations, the use of AI in skill assessments is a rapidly growing and promising field of research.

Motion/Contact Vibrations

Many groups across all surgical platforms are looking for methods of assessment that use purely objective psychometrics to eliminate the inherent bias of human judges. In our review, Gomez and colleagues(Gomez et al. 2015) had some success using contact vibration as a surrogate for robotic skill in a series of dry-lab tasks. Their study demonstrated that lower vibration and force-derived metrics were recorded in their cohort of experienced robotic surgeons as compared to novices. This novel evaluation method showed good construct validity in 10 out of 15 metric-task correlations, demonstrating that this purely objective method has utility in formative skill assessments. However, this and similar unidimensional psychomotor assessments may not reflect the full competence, or lack thereof, and must demonstrate correlation with patient outcomes before they are accepted on the main stage of surgical assessment.

Armrest Load

Two studies from Yang et al.(Yang, Perez, et al. 2017; Yang, Zhen, et al. 2017) quantified armrest load and surgeon ergonomics as methods of both assessment and educational intervention in robotic surgery training. They found they could distinguish between surgeons with different robotic experience in a simulated environment, as well as shorten the simulation-based learning curve of novice trainees by building in a real-time feedback mechanism that alerts the trainee about excessive weight applied to the console armrest. This metric has potential as a means of both improving trainee acquisition of technical competency and complementing assessments of surgeon skill in training curricula.

Literature Quality Assessment

The mean MERSQI score for all included articles was 12.8, which falls short of the 14/18 mark that indicates ‘high quality’. Articles found to have a score of 14 or higher are detailed in Table-4.

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
Vlaovic et al. (2008)	101 T	Dry, TS	5-day laparoscopic training program. Includes 2-3 hrs of lectures, daily practice on pelvic trainers and VR simulators, and training on porcine models and human cadavers. Assessed ring transfer, suture threading, cutting, and suturing by expert examiner	OSATS	Post-course robotic performance was significantly improved ($p < 0.001$)	14
Davis et al. (2010)	3 R, 4 F	OR, TS	Standardized method of evaluating performance in robot-assisted radical prostatectomy using time, autonomy scale and end-product assessment by expert surgeons.	Time, quality of results relative to staff, short term patient outcomes	Time to completion was longer for trainee's vs staff ($p < 0.001$), basic vs advanced tissue dissection and suturing. No increase in adverse short-term outcomes was observed	14
Kumar et al. (2012)	6 novice, 2 expert	Dry, TS	Support Vector Machines (SVM) to classify expert-novice operational skills. Assessed master workspace adjustment, camera manipulation skills, unsafe motion and collisions by computer for	Support Vector Machines (SVM)	Model correctly classified 91.7% for master workspace and 88.2% for camera manipulation	14

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
			manipulation, suturing, transection, and dissection			
Foell et al. (2013)	29 R, 16 F, 8 S	VR/Dry, TS	Participants included urology, obstetrics and gynecology, and thoracic surgery. Assessed Camera Targeting 1, Peg Board 1, Match Board 1, Thread the Rings, Suture Sponge 1, Ring Walk 2, and Peg Board 2 by dVSS, and compared to dry-lab performance on robotic console	dVSS metrics, time/number of errors	Performance on dVSS modules had moderate-strong correlation with time/error assessment on robotic console in dry-lab setting	14
Yamany et al. (2015)	13 R	Dry, TS	Effect of 24-hr call on suturing performance of residents with or without prior robotic simulator experience. Participants included urology and general surgery. Assessed time to completion of exercise,	dVSS metrics	Time to completion, needle loading, and knot tying were significantly increased postcall ($p < 0.05$). Prior simulator experience did not have significant benefits in	14

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
			needle loading, knot tying by dVSS		postcall performance ($p < 0.05$)	
Whitehurst et al. (2015)	7 R, 8 F, 5 S	dV-Trainer/Dry/Wet (swine), TS	Compared robotic performance between training in a VR or dry lab setting. Participants included gynecology, urogynecology, gynecologic oncology, reproductive endocrinology, and urology. Assessed cystotomy closure on swine model by blinded expert surgeons	dV-Trainer metrics, GEARS	Training modalities did not differ significantly: 2.83 ± 0.66 for VR cohort, 2.96 ± 0.77 for dry cohort, $p = 0.690$	14
McVey et al. (2016)	11 R, 21 F	Box-Trainer, TS	Effect of baseline laparoscopic skill on robotic skill before and after robotic surgery basic skills training course. Participants included urology, gynecology, thoracic surgery, and	Time, number of errors	Baseline laparoscopic intracorporeal suturing and knot tying (ISKT) performance strongly correlated with robotic performance	14

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
			general surgery. Assessed by two blinded content experts using Likert scale global rating score		(p = 0.01 for peg transfer, p < 0.01 for ISKT). IRR = 0.9	
Chowriappa et al. (2013)	15 novice, 12 expert	VR, TS	Assessed fourth arm control, coordinated tool control, ball placement, and needle handling and exchange by RoSS simulator	RSA-Score	Expert cohort performed significantly across all tasks: p = 0.002 for fourth arm control, p < 0.001 for coordinated tool control, p < 0.001 for ball placement, p < 0.001 for needle handling and exchange	14.5
Hung et al. (2015)	15 novice, 13 intermediate, 14 expert	AR/VR, TS	Developed simulation platform for robotic partial nephrectomy. Includes augmented reality content and virtual reality renorrhaphy. Assessed by blinded expert reviewer	dV-Trainer metrics, GEARS	Simulation platform demonstrated strong face, content, and construct validity. Virtual reality renorrhaphy performance correlated significantly with porcine model (r = 0.8, p < 0.0001)	14.5

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
Schommer et al. (2017)	34 R	dV-Trainer, TS	Compared access to robotic technology to robotic skill between residents attending a training course in 2012 and 2015. Assessed Camera Targeting 2, Energy Dissection 1, Needle Targeting, and Peg Board 1 by dV-Trainer	dV-Trainer metrics	Robotic performance was significantly better in the 2015 cohort than 2012 ($p < 0.001$). Access to robot console correlated with better scores in Camera Targeting 2 ($p = 0.02$) and Peg Board ($p = 0.04$)	14.5
Raison et al. (2017)	102 R, 121 S	dV-Trainer, TS	Set benchmark scores to achieve competency in robot skills. Assessed basic (Pick and Place, Camera Targeting 1, Peg Board 1) and advanced (Thread the Rings 1, Suture Sponge) tasks by dV-Trainer	dV-Trainer metrics	Using a benchmark score of 75% of the mean expert score, novice trainees achieved competency in basic but not advanced tasks. Intermediate trainnes achieved competency in basic tasks and Suture Sponge	14.5

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
Song Xu et al. (2016)	11 Robotic-experienced, 7 Laparoscopic-Experienced, 9 Control	Xperience Team-Trainer (XTT)	Establish initial validity evidence for a team-based robotic surgery simulator, including bedside assistant involvement. Evaluated simulation performance as assistant and console surgeon using the XTT.	XTT Metrics Modified GOALS	Demonstrated that scores on XTT correlate with both robotic experience and performance on the console. The robotic and laparoscopic experienced surgeons outperformed controls in all exercises.	14.5
Stegemann et al. (2013)	53 participants; 9 medical students, 26 residents, 10 fellows, and 8 practicing surgeons	Box trainer, TS	Provide validity evidence, demonstrating that Fundamental Skills of Robotic Surgery (FSRS) curriculum completion improves performance on tasks completed with actual daVinci console in simulation setting.	Number of errors, camera/clutch use	Although no differences between study arms, control group showed significant improvement from baseline on repeat daVinci console scores when allowed to crossover into FSRS arm	14.5

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
Lendvay et al. (2013)	27 R, 24 S	VR/Dry, TS	Effect of VR warm-up on robotic performance in similar and dissimilar tasks. Participants included general surgery, urology, and gynecology. Assessed rotating rocking pegboard and intracorporeal suturing by computer	Time, cognitive and technical errors, tool path length, economy of motion	Warm-up cohort performed significantly better in time ($p = 0.001$) and path length ($p = 0.014$) for similar tasks (rotating rocking pegboard) and significantly better in global technical errors ($p = 0.020$) for dissimilar tasks (intracorporeal suturing)	15

Table 4 Description of high quality evidence (MERSQI ≥ 14)

Study	Trainees	Setting, Type of Assessment	Assessment summary	Measurement Tool	Conclusion	MERSQI
Tarr et al. (2014)	99 R	Dry, TS	Compared robotic performance before and after an unstructured or structured robotic training curriculum. Structured curriculum included specific instructions and goal times to achieve before proceeding to the next task. Participants included gynecology and urology. Assessed manipulation, transection, knot tying, and suturing by expert examiner	OSATS	Structured cohort performed significantly better in transection ($p < 0.05$), while unstructured cohort performed significantly better in knot tying ($p < 0.05$). No significant differences were observed in manipulation and suturing	15

10.5.5.1.4 Discussion

This review has highlighted the various assessment methods that exist in evaluating technical skill when performing robotic surgery in urology. This area of research is still actively evolving, and while this article has summarized the methods used to date, we expect that applications and diversity of these instruments will continue to expand and develop as the paradigm of competency-based training becomes the standard.

We have outlined the various efforts made in assessing technical skill in urologic robotic surgery, and while the literature is diverse, we have shown some homogeneity in the underlying principles of assessment being employed. As in most studies assessing technical skill, global rating scales continue to be more popular than task-specific checklists, due to their broader applicability and ease of use(Regehr et al. 1998).

Although many of these assessment tools can be applied across all types of robotic surgery, urology will likely lead the movement toward the use of these assessments in surgeon accreditation, as opposed to its current place in the formative setting only. Educators and licensing stakeholders will pay attention in urology especially, as the role of surgeon performance in patient safety and outcomes continues to be investigated in this space(M. G. Goldenberg, et al. 2017). This emerging evidence will likely lead to the incorporation of assessments of technical and non-technical skill into licensing practices at a local or national level(J. Y. Lee et al. 2011). As of now, the accreditation process remains under the sole control of the hospitals(Zorn et al. 2009), and there is no established use of summative technical skill assessments in robotic surgery for the purposes of credentialing.

There are specific limitations of this review and the included research presented. A major issue that is prevalent throughout the robotic assessment literature is the comparison of novice and expert surgeons as a source of validity evidence. In order to frame an assessment in a specific context, i.e. low-stakes vs. high-stakes, it is crucial that the assessment construct be clearly defined. Making decisions of competency

within a training program requires the chosen assessment to distinguish between trainees who have met a predefined set of criteria from those that require further remediation. In contrast, an assessment designed for credentialing robotic surgeons after training must be able to distinguish between those who will have satisfactory patient safety and clinical outcomes and those that do not. Unfortunately, much of the literature choses to compare groups at the extremes of skill to allow for highly statistically significant differences in 'scores' between cohorts. Secondly, it is important to note that the *internal structure* and *response process* validity for simulators is often hard to quantify. Although computer-generated and algorithm-based scoring metrics are assumed to be accurate and reliable, it is still essential that manufacturers and academics strive to provide this validity evidence as robustly as possible, by clearly describing how their scoring components are tabulated and weighted, and any quality control process that are undertaken in the development of scoring algorithms.

Importantly, most studies in this review contribute at least one source of validity evidence for their described assessment tool, as shown in Table-1. However, gaps in the supporting evidence are present in the majority of these studies, and emphasis should be placed moving forward on addressing this. Despite all studies contributing one or more source of validity evidence for a given assessment, many various data elements that make up each of Messick's five domains of validity were vastly underrepresented.(Cook et al. 2014) Of note, *internal structure* and *response process* evidence was fairly homogenous in nature across the included literature. While interrater reliability statistics were more commonly reported, other important *internal structure* data such as internal consistency (reliability across the domains of the assessment tool) and test-retest reliability (reliability across different sittings or versions of the assessment) were rarely included or described in these studies. Additionally, crucial components of *response process* evidence such as rater data analysis (understanding rater disagreements or inconsistencies) and effects of rater training (comparison of scores between trained and untrained raters) were also not addressed by most of these studies. Typically, *response process* evidence in these studies consisted only of descriptions of rater training, and the use of video capture to ensure

quality control of testing data. These gaps in evidence may reflect the investigator's use of outdated taxonomies of validity when designing these studies, including decisions around the type of data to calculate and report in their manuscript.

Recommendations

Using Messick's Conceptual Framework of Validity(Messick 1994), we have systematically gathered and quantified the validity evidence supporting technical and computer-based VR assessments of robotic surgical skill, to provide evidence-based recommendations on how best to implement these assessment tools in postgraduate training and, in future, credentialing practices.

It is clear from our review that assessments of technical skill using the GEARS metric are strongly supported with robust validity evidence in a wide range of settings, from ranking medical students in the residency match to distinguishing 'high' and 'low' performances of a single, high-volume surgeon. It provides reliable ratings of trainee or faculty performance in real-time assessments in the lab or operating room, or when used in video-based evaluation by expert raters or laypeople through crowdsourcing. However, it is important to note that while many studies report a high to very-high interrater reliability, this is not true of all the included literature. We must stress to educators the importance of training faculty in the use of these assessment rubrics, and early identification of raters who are outliers in their scoring of trainee technical skill. Another option for technical skill evaluation is the OSATS tool, long seen as a gold-standard amongst GRS assessments. This scale has been used in multiple settings in the literature, nd has an excellent evidence-base when applied in all testing environments, including dry lab, simulation/VR, and the operating room. Its broadly applicable domains allow it to be used and easily compared with assessments in open and laparoscopic surgery, making it an attractive option for evaluating technical competency across multiple surgical platforms.

It is difficult to provide a single recommendation on computer-based VR assessment, but the validity evidence for both the dVSS and the dV-Trainer systems in low-stakes assessments is strong. Both platforms have been shown to distinguish between trainees and surgeons of differing skill levels, and both have demonstrated *response process* validity through test-retest methodology and correlation of computer-generated scores with human ratings. Like the GEARS score, these platforms can be used in the training and assessment of participants with a range of robotic surgery experience, but most of the literature supports use in postgraduate education rather than in high-stakes assessments, such as credentialing, as evidence of their ability to predict clinical outcomes is currently lacking.

10.5.5.1.5 Conclusion

As the competency-based education model of surgical training continues to become more universal(J. R. P. I. MD 2016; Canada 2014; Hammond et al. 2005), it is imperative that educators understand not only the milestones set forth by their governing bodies, but also the methods in which these milestones are defined. We have provided a summary of the current literature describing technical skill assessments in urological robotic surgery, and provide evidence-based recommendations of how one may implement these into a competency-based curriculum. Competency in surgical skill must be defined by content experts, through objective means, and the validity evidence of the assessment tools discussed here should give educational stake-holders confidence in making judgments on their trainee's ability. Despite this, the question of how to best create summative assessments of surgical skill remains unanswered. As demonstrated in this review, there are efforts on multiple fronts, from the simulation lab to the operating room.

10.6 Setting Performance Standards in Technical Skill

10.6.1 Systematic Review to Establish Absolute Standards for Technical Performance in Surgery

10.6.1.1 Introduction

Training and accreditation in surgery is a shifting landscape(J. R. P. I. MD 2016). Medical education has been moving away from traditional time-dependent learning in favour of competency-based learning. Recently, regulating bodies such as the Accreditation Council for Graduate Medical Education(Kohlwes et al. 2011) and the Royal College of Physicians and Surgeons of Canada(Frank et al. 2015) have begun incorporating competency-based learning in their curricula. In Canada, the Canadian Medical Education Directives for Specialists (CanMEDS) 2015 framework(Frank et al. 2015) encompasses seven domains, in which all physicians training in Canada must demonstrate competency at the time of certification. In any training system, there must be clear processes for both formative (low-stakes) assessments that give trainees ongoing evaluation and feedback and summative (high-stakes) assessments that can distinguish between satisfactory and remedial levels of performance, skill or knowledge(Cendan et al. 2013; Epstein 2007). These changing principles are accompanied by a demand for methods of assessment that distinguish between those who are competent and those who are not. Standard setting allows stakeholders to determine competence and non-competence by setting cut-off scores or benchmarks(Cendan et al. 2013).

The historical uses of standard-setting methodology in medicine have been mainly for written assessments at the undergraduate level(Kaufman et al. 2000) and at certification following postgraduate training(Norcini 2003). Its application to procedural assessments is a more recent use of this methodology(Cendan et al. 2013).

Depending on the nature of an assessment, this standard needs to be built around either the difficulty of the test or the performance of those taking the test(Norcini 2003). These two concepts are referred to as item-centred and participant-centred

respectively. The key aspect differentiating these two categories is whether expert judges have yet observed the performance of the examinees or not. Making participant-centred judgements based on the cohort's performance on the test or skill allows experts to make independent assessments of individuals performing the test, whereas item-centred methods require judges to conceptualize the borderline student and then make judgements of how this theoretical student would perform on the test itself(Sturmberg & Hinchy 2010).

Item-centred and participant-centred are both examples of absolute standards, also referred to as criterion-referenced standards(Downing et al. 2010). These are in contrast to relative or norm-referenced standards, which establish a pass mark by comparing the test group with another defined, often expert, group(Downing et al. 2010). Conversely, absolute standards utilize content experts to set a more objective pass mark that is representative of the task and not simply a reflection of the performance of another group(Downing et al. 2010). It is important that summative assessments are developed in a careful and thoughtful way, as they may have a meaningful impact on students and stakeholders alike in a given training system(Szasz et al. 2014). In addition to the use of criterion-referenced standards for procedural assessment, these standards must also be valid and reliable, with a clear purpose that fits the objectives of the curriculum/assessment of which they are a part.

The objective of this systematic review was to identify absolute standard-setting methodologies that set technical performance benchmarks with regard to procedural technical skills. Additionally, the review aimed to assess the quality of the included studies, using the Medical Education Research Quality Index (MERSQI)(Reed et al. 2007).

10.6.1.2 Methods

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement(Liberati et al. 2009) was utilized.

Eligibility criteria

Studies eligible for inclusion were those evaluating medical students, interns, residents, fellows or staff physicians/surgeons in any medical or surgical specialty that participated in an assessment of procedural skill wherein a performance standard was set using an absolute methodology. Studies were eligible regardless of publication status (full-text manuscript, conference abstract). RCTs and observational studies, including cohort, case–control, case series and cross-sectional studies, were all eligible for inclusion.

Information sources

One author and a librarian at St Michael's Hospital Library in Toronto, Canada, conducted independent searches in Ovid MEDLINE, Embase Classic, PsycINFO and the Cochrane Library.

Search

Medical subject headings (MeSH) terms used in the search included "Professional Competence", "Clinical Competence", "Credentialing", "Accreditation", "Licensing", "Surgical Procedures", "Psychomotor Performance", "Professional Standard", "Physician" and "Medical Student". Keywords used included "surgeon", "clinical skill", "clinical performance", "proficiency", "competence", "technical", "pass/fail" and "cutoff". Included among the keywords of the search were the names of absolute methods of standard setting, for example "Angoff", Borderline" and "Contrasting Groups".

Titles of articles resulting from the search and corresponding abstracts were reviewed initially and articles eligible for full-text review were identified. These articles

were then analysed further to ensure no articles referenced therein were missed for inclusion in the full-text review. Duplicates were identified and removed.

Study selection (Inclusion and Exclusion)

Any study in the medical or surgical literature that used absolute standard-setting methodology to establish a performance standard in technical skill, including surgical, endoscopic and bedside procedures, involving original research and described in English were included.

Opinion letters, reviews, case reports, letters to the editor and non-original research articles were excluded. Abstracts and conference presentations were excluded only if the corresponding published article was captured in the search. Also excluded were studies that assessed non-procedural skills, such as cardiorespiratory resuscitation and other acute-care management scenarios. Finally, studies using only relative standards were excluded, as they were considered to be outside the scope of this review.

Two authors reviewed articles independently, and any differences were resolved by consensus.

Data collection process

Data were collected from the included studies in a methodical manner, and included field of study, standard-setting methodology employed, study design, type of participants, and type of outcome assessed.

Quality Assessment

The MERSQI was used by two authors to assess study quality. The MERSQI includes eight domains: study design, institutions sampled, response rate, type of data, validity

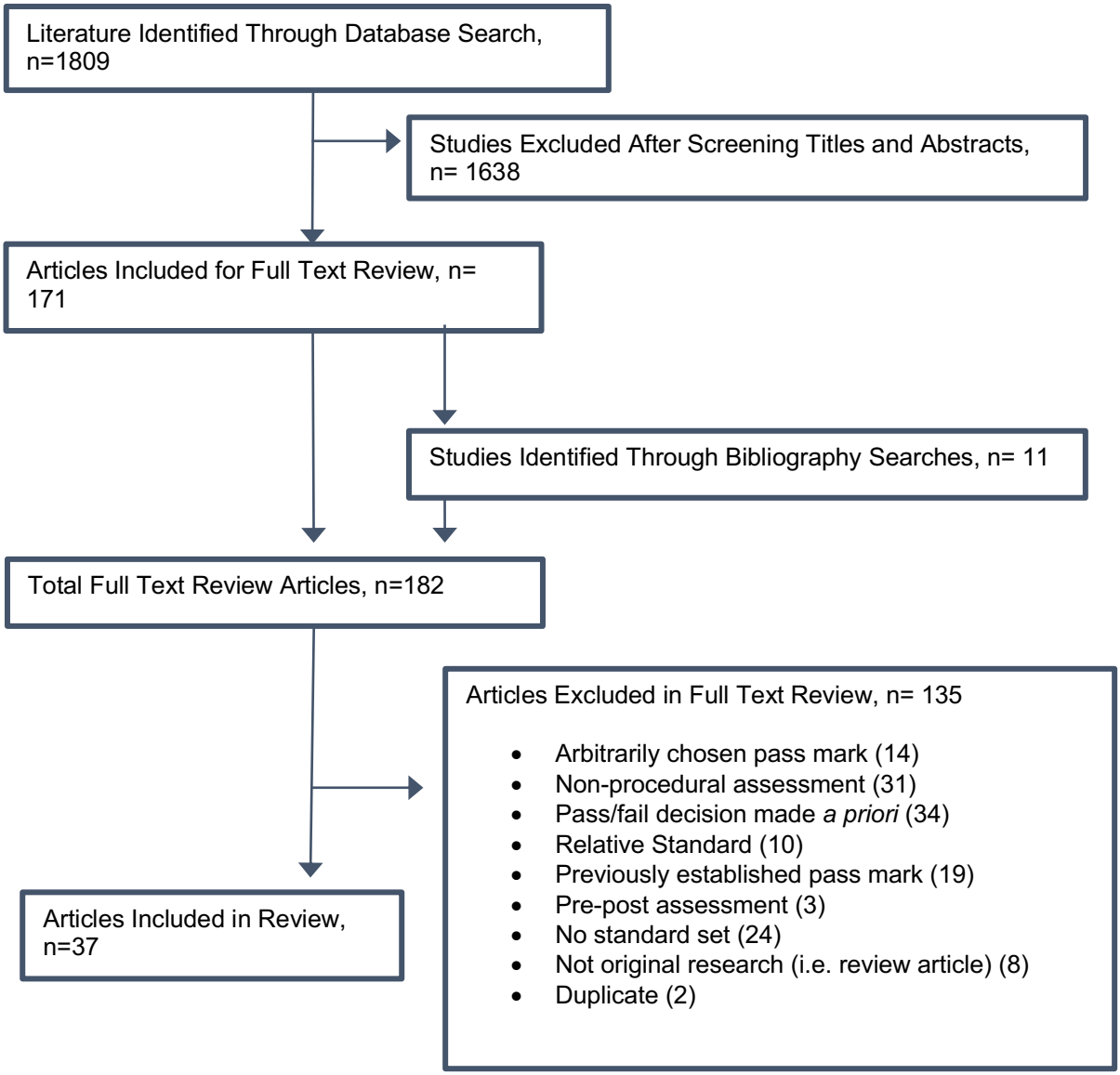
evidence for evaluation of instrument scores, sophistication of data analysis, appropriateness of data analysis, and outcome. It uses the principles of Messick's conceptual framework to assess methodological validity(Messick 1994). A score of 14 of 18 was taken as the minimum to be considered a high-quality study, as used elsewhere(Lin et al. 2016).

10.6.1.3 Results

Included studies

The search yielded 1809 studies initially. Titles and abstracts of these studies were screened in a systematic fashion, and decisions of inclusion or exclusion were made according to the predetermined criteria described above. This process yielded 171 articles for full-text review. After 11 more papers were included through searching relevant bibliographies of included studies, the full-text review included 182 articles. Following full-text review, 37 articles met the inclusion criteria and were included in the systematic review. Reasons for exclusion are described in *Fig. 1*.

Figure-2: PRISMA Flow Chart



Participants

Standard-setting studies incorporated participants at various stages of their career. Of the 37 articles, 26(Walzak et al. 2015; Thinggaard et al. 2015; Jacobsen et al. 2015; Tolsgaard et al. 2014; Tjiam et al. 2012; Cohen et al. 2013; Konge et al. 2012; Wayne et al. 2007; Jelovsek et al. 2010; Huang et al. 2009; McCluney et al. 2007; Beard 2005; Fraser et al. 2003; I. C. Green et al. 2013; Stefanidis et al. 2006; Diwadkar et al. 2009; King et al. 2015; Wayne et al. 2008; Kowalewski et al. 2015; de Montbrun et al. 2016; Barsuk, Cohen, Vozenilek, et al. 2012; Barsuk, Cohen, Caprio, et al. 2012; Burch et al. 2005; N. N. MD et al. 2015; E. N. T. MD et al. 2014; de Montbrun et al. 2015) included postgraduate trainees (residents and interns). For example, Tjiam and colleagues(Tjiam et al. 2012) established certification standards for residents in laparoscopic urological skills by comparing global ratings of a cohort of staff surgeons and trainees stratified by case volume into expert, intermediate and novice groups. Seventeen articles(Thinggaard et al. 2015; Jacobsen et al. 2015; Tolsgaard et al. 2014; Tjiam et al. 2012; Konge et al. 2012; McCluney et al. 2007; Beard 2005; Fraser et al. 2003; Kowalewski et al. 2015; N. N. MD et al. 2015; Thomsen et al. 2015; Preisler et al. 2015; Pedersen et al. 2014; Kissin et al. 2013; Konge et al. 2013; Vassiliou et al. 2013; Svendsen 2014) used fully qualified surgeons and physicians. Pedersen and co-workers(Pedersen et al. 2014) used expert orthopaedic surgeons as the competent group on a hip fracture simulator, and set standards that trainees must reach before progressing to the clinical environment. Fellows are a unique group of participants, as they can be included under the trainee umbrella, and yet are often experienced operators at procedures tested in standard-setting exercises. In this review, six articles(McCluney et al. 2007; Preisler et al. 2015; Kissin et al. 2013; Sedlack 2011; Barsuk et al. 2009; Sedlack & Coyle 2016) specified the use of fellows as study participants, and two articles included medical student participants(Yudkowsky et al. 2014; Fraser et al. 2003).

Participant-centred methodology

Of the 37 included articles, 24 used participant-centred methodology to set standards in technical performance. Of these, 18(Thinggaard et al. 2015; Jacobsen et al. 2015; Konge et al. 2012; Tjiam et al. 2012; McCluney et al. 2007; Fraser et al. 2003; Stefanidis et al. 2006; King et al. 2015; Kowalewski et al. 2015; de Montbrun et al. 2016; N. N. MD et al. 2015; Thomsen et al. 2015; Preisler et al. 2015; Pedersen et al. 2014; Konge et al. 2013; Vassiliou et al. 2013; Svendsen 2014; de Montbrun et al. 2015) took place in a simulated setting, and six(Tolsgaard et al. 2014; Beard 2005; Diwadkar et al. 2009; Kissin et al. 2013; Sedlack 2011; Sedlack & Coyle 2016) in a clinical setting. The type of assessment tools used in the standard-setting exercise are shown in *Table-5*. Assessment tools used either a global rating or task-specific metric, or both. Seven studies(Jacobsen et al. 2015; Tolsgaard et al. 2014; McCluney et al. 2007; Kowalewski et al. 2015; Thomsen et al. 2015; Kissin et al. 2013; de Montbrun et al. 2015) that employed participant-centred methods used only a global rating system, nine used only task-specific checklists/simulator-derived metrics(Stefanidis et al. 2006; Pedersen et al. 2014; Konge et al. 2013; Vassiliou et al. 2013; N. N. MD et al. 2015; Svendsen 2014) or task-specific Likert scale assessments(Thinggaard et al. 2015; Tjiam et al. 2012; Fraser et al. 2003), and eight(Diwadkar et al. 2009; King et al. 2015; de Montbrun et al. 2016; Preisler et al. 2015; Sedlack 2011; Sedlack & Coyle 2016; Konge et al. 2012; Beard 2005) used both in combination to set standards. de Montbrun *et al*.(de Montbrun et al. 2016) created a national colorectal surgery Objective Structured Clinical Examination (OSCE), in which participants were scored on their performance using a combination of a task-specific checklist and overall global rating on a 5-point Likert scale. McCluney and colleagues(McCluney et al. 2007) set standards in simulated laparoscopic skill using their novel global rating method, the Global Operative Assessment of Laparoscopic Skills (GOALS), and demonstrated its ability to predict intraoperative performance of trainees.

Table-5 Methods and Location of Standard Setting

Standard Setting Method	Assessment Setting		Type of Assessment Used			Total Number of Studies
	Simulation	Clinical	GR	TS	Both	
Participant-Centered	18	6	7	9	8	24
Contrasting-Groups	*12*	*5*	*4*	*6*	*7*	*18*
Borderline-Group	*1 (1)†*	*1*	*1*	*0*	*1*	*2*
Generalized Examinee-Centered	*1*	*0*	*0*	*1*	*0*	*1*
Receiver Operator Characteristic Curve (ROC)	*4 (1)†*	*0*	*2*	*2*	*0*	*4*
Item-Centered	12	1	1	10	2	13
Angoff	*12 (1)†*	*1*	*1*	*10*	*2*	*13*
(+ Hofstee††)	*(6)*	*(0)*	*0*	*6*	*0*	*(6)*
(+ Ebel††)	*(1)*	*(0)*	*0*	*0*	*1*	*(1)*
Total	**30**	**7**	**8**	**19**	**11**	**37**

† *Used as a secondary method of standard setting*
††*Hofstee and Ebel used in conjunction (or for comparison) with Angoff method*
GR = Global Rating, TS = Task-Specific

Three articles that used participant-centred methodology included a second standard-setting method, either as a way of directly comparing the standard created (de Montbrun et al. 2015; Vassiliou et al. 2013), or to set standards for a written component of the assessment(Konge et al. 2012) (*Tables 5* and *6)*.

Item-centred methodology

Item-centred standard-setting methods were used 13 times. Of these, 12(Wayne et al. 2007; I. C. Green et al. 2013; Wayne et al. 2008; Barsuk, Cohen, Caprio, et al. 2012; Burch et al. 2005; Barsuk et al. 2009; Walzak et al. 2015; Yudkowsky et al. 2014; E. N. T. MD et al. 2014; Barsuk, Cohen, Vozenilek, et al. 2012; Huang et al. 2009; Cohen et al. 2013) were derived from a simulated surgical environment and only one(Jelovsek et al. 2010) from a clinical setting (*Table 5*). Unlike the participant-centred literature, the majority of item-centred studies used task-specific metrics to set standards, ten in total(Cohen et al. 2013; Wayne et al. 2007; Huang et al. 2009; Wayne et al. 2008; Barsuk, Cohen, Caprio, et al. 2012; Burch et al. 2005; E. N. T. MD et al. 2014; Barsuk et al. 2009; Yudkowsky et al. 2014; Barsuk, Cohen, Vozenilek, et al. 2012), whereas only one used global rating alone(I. C. Green et al. 2013), and two(Walzak et al. 2015; Jelovsek et al. 2010) used a combination of global rating and task-specific methods (*Table 5*. One group of authors has published multiple studies(Wayne et al. 2007; Wayne et al. 2008; Barsuk, Cohen, Caprio, et al. 2012; Barsuk et al. 2009; Barsuk, Cohen, Vozenilek, et al. 2012) that utilize a task-specific checklist to assess internal medicine residents' technical skill in a variety of bedside procedures. These studies were able to use Angoff and Hofstee methods (both item-based) to set standards in simulated tasks, as the environment and steps of the procedure were highly controlled. Conversely, Green and co-workers(I. C. Green et al. 2013) developed an adaption of the objective structured assessment of technical skill (OSATS)(Martin et al. 1997), the Robotic-OSATS (R-OSATS), to assess trainees in robotic surgery as they progressed through a highly structured simulator curriculum.

Table-6 Standard Setting Method and Type of Procedure Assessed

Standard Setting Method	**Type of Procedure Assessed**			
	Open Surgery†	**Endoscopic/ Arthroscopic**	**Laparoscopic/ Robotic**	**Bedside Procedures**
Participant-Centred	6	9	7	2
Contrasting-Groups	*4*	*9*	*3*	*1*
Borderline-Group	*1*	*0*	*0*	*1*
Generalized Examinee-Centred	*0*	*0*	*1*	*0*
Receiver Operator Characteristic Curve (ROC)	*1*	*0*	*3*	*0*
Item-Centred	0	0	3	10
Angoff	*0*	*0*	*3*	*10*
(+ Hofstee)*	*0*	*0*	*0*	*6*
(+ Ebel)*	*0*	*0*	*0*	*1*
Total	**6**	**9**	**10**	**12**

† Two studies assessed a combination of open and laparoscopic surgical skills, and are included here [33,49]

** Hofstee and Ebel used in conjunction (or for comparison) with Angoff method*

All articles included in the studies that used item-centered methodology employed the Angoff method as the standard setting technique, although six of these studies(Cohen et al. 2013; Wayne et al. 2007; Wayne et al. 2008; Barsuk, Cohen, Caprio, et al. 2012; Barsuk et al. 2009; Barsuk, Cohen, Vozenilek, et al. 2012) also used the Hofstee method, to create a comparative standard. One other paper(Walzak et al. 2015) used the Ebel method. These additional methods are indicated in *Table 5*.

Types of Procedure Assessed

The studies included created benchmarks across a variety of procedures, in both the simulated and clinical settings (*Table 6*). Six studies(Thomsen et al. 2015; Pedersen et al. 2014; Jelovsek et al. 2010; Beard 2005; de Montbrun et al. 2015; de Montbrun et al. 2016) assessed open surgical procedural skills, two(de Montbrun et al. 2015; de Montbrun et al. 2016) of which included evaluation of a small number of laparoscopic skills. Nine studies(Jacobsen et al. 2015; Preisler et al. 2015; Konge et al. 2013; Konge et al. 2012; Sedlack 2011; Vassiliou et al. 2013; Sedlack & Coyle 2016; N. N. MD et al. 2015; Svendsen 2014) evaluated endoscopic or arthroscopic surgery. As noted in *Table 6*, all open and endoscopic procedure studies used participant-centred methodology. Ten studies(Thinggaard et al. 2015; Tjiam et al. 2012; McCluney et al. 2007; Fraser et al. 2003; I. C. Green et al. 2013; Stefanidis et al. 2006; Diwadkar et al. 2009; King et al. 2015; E. N. T. MD et al. 2014; Kowalewski et al. 2015) determined competency benchmarks in laparoscopic and robotic surgery, ten(Walzak et al. 2015; Yudkowsky et al. 2014; Cohen et al. 2013; Wayne et al. 2008; Huang et al. 2009; Barsuk et al. 2009; Wayne et al. 2007; Barsuk, Cohen, Vozenilek, et al. 2012; Burch et al. 2005; Barsuk, Cohen, Caprio, et al. 2012) set standards in bedside procedures using item-centred methodology, and only two(Tolsgaard et al. 2014; Kissin et al. 2013) used participant-centred techniques.

Standard-Setting Judges

Included studies used judges who were selected to create the standards in technical skill. *Table 7* shows the categories of judge used, ranging from experts described as being trained in standard-setting methodology to computer based, simulator-derived scores. Participant-centred studies used expert judges trained in standard setting in eight studies(Thinggaard et al. 2015; Tolsgaard et al. 2014; Tjiam et al. 2012; Konge et al. 2012; McCluney et al. 2007; Diwadkar et al. 2009; King et al. 2015),[34], as did item-centred studies(Cohen et al. 2013; Wayne et al. 2008; Barsuk, Cohen, Caprio, et al. 2012; Barsuk et al. 2009; Walzak et al. 2015; E. N. T. MD et al. 2014; Barsuk, Cohen, Vozenilek, et al. 2012; Jelovsek et al. 2010). Four further participant-centred(Preisler et al. 2015; Sedlack & Coyle 2016; Kissin et al. 2013; Beard 2005) and four item-centred(Wayne et al. 2007; Huang et al. 2009; Burch et al. 2005; Yudkowsky et al. 2014) studies used content expert judges, although it was not explicitly stated that these judges were trained in standard setting. As expected, item-centred methods did not use simulator-generated metrics to set standards, but six participant-centred studies(Jacobsen et al. 2015; Pedersen et al. 2014; Konge et al. 2013; Vassiliou et al. 2013; N. N. MD et al. 2015; Svendsen 2014) employed these as the basis for forming standards, usually based on surgeon experience as the method of dividing the cohort into comparable groups.

Table-7 Judges Used in Absolute Standard Setting

Absolute Methods	Number of Studies
Participant-Centred	n=24 (%)
Content Experts, Trained in Standard Setting	*8 (33)*
Content Experts, Untrained in Standard Setting	*4 (17)*
Non-Experts, Trained in Standard Setting	*1 (4)*
VR Simulator Generated Metrics	*6 (25)*
Not Specified In Article	*5 (21)*
Item-Centred	n=13 (%)
Content Experts, Trained in Standard Setting	*8 (62)*
Content Experts, Untrained in Standard Setting	*4 (31)*
Non-Experts, Trained in Standard Setting	*0 (0)*
VR Simulator Generated Metrics	*0 (0)*
Not Specified In Article	*1 (7)*

**One study uses both contrasting groups AND Hofstee methods*

Methodological Quality

Overall study quality using the MERSQI tool is summarized in *Table 8*. Of the 37 articles, 17 met the benchmark. The mean quality score of all 37 articles was 13.67 of 18. The mean quality of participant-centred standard-setting studies was higher than that of item-centred ones (*Table 8)*.

Table 9 illustrates the specific types of procedure assessed, divided into medical, surgical, and obstetrics and gynaecology specialties along with MERSQI scores. Some studies included participants from more than one specialty, two(Vassiliou et al. 2013; Svendsen 2014) assessing colonoscopy skill and one(Thinggaard et al. 2015) in laparoscopy. One study(Konge et al. 2012) evaluated its participants performing a combination of clinical and simulated procedures.

Table-8 Quality Assessment of Methodology Using the Medical Education Research Quality Index (MERSQI)

Absolute Method	Mean MERSQI Score (/18) (Range)
Participant-Centred (n)	13.91 (12.5-15.5)
Contrasting-Groups (17)	*13.90*
Borderline Group (2)	*14.25*
Generalized Examinee-Centred (1)	*14.5*
Receiver-Operator Characteristic Curve (ROC) (4)	*13.63*
Item-Centred (n)	13.17 (11-14.5)
Angoff (13)	*13.17*
Hofstee (6)	*13.41*
Ebel (1)	*14.00*
Total (37)	*13.70 (11-15.5)*

**MERSQI Score >14 is considered "high-quality" evidence*

Table-9 Specific Procedures Assessed in Included Literature (With corresponding MERSQI Score)

Study Setting	Medicine	MERSQI Score(s)	Surgery (n)	MERSQI\ Score(s)	Obstetrics and Gynecology (n)	MERSQI Score(s)
Clinical (Real World)	Musculoskeletal Ultrasound Colonoscopy *Bronchoscopy	 15 12.5, 12.5 13.5	Saphofemoral Disconnection	 14.5	Obstetrical Ultrasound Vaginal Hysterectomy Laparoscopic Hysterectomy	14.5 13.5 12.5
Simulation	Thoracentesis Paracentesis Endobronchial Ultrasound *Bronchoscopy Basic Bedside Procedures *(Phlebotomy, Intubation, etc.)* Vascular Line Insertion Lumbar Puncture Colonoscopy	14.5 13.5 15.5 13.5 14, 12.5, 11 15, 12.5, 12, 12 13.5 15, 14.5, 13.5, 12.5	Basic Surgical Skills *(Suturing, Foley Catheter etc.)* Cataract Surgery Hip Fracture Knee Arthroscopy Colonoscopy† Laparoscopy Robotic Surgery	13.5, 12.5 5.5 15 15 15, 13.5 14.5, 14.5, 14, 14, 14, 13.5 12.5	Vaginal Surgery Robotic Surgery Laparoscopic Surgery ††	13.5 13.5 14, 13.5

†Two studies assessing colonoscopy skill used surgeon-participants
††One study used both surgical and OBGYN participants
**Study used both simulation and clinically obtained video footage in assessmen*

10.6.1.4 Discussion

There has been a clear evolution in the use of benchmark assessments that reflects changes in the field of medical education. Most recently, this competency-based training or competency-by-design philosophy has caused the primary use of standard setting to include recognition of competency(Beard 2005). This systematic review of standard setting used in technical procedural skill assessment has confirmed that this methodology could satisfy the growing need for competency-based summative assessments in medicine and surgery(Barsuk, Cohen, Vozenilek, et al. 2012; Jelovsek et al. 2010; Beard et al. 2011).

The most commonly established standards identified used the participant-centred contrasting-groups method(Thinggaard et al. 2015; Jacobsen et al. 2015; Tolsgaard et al. 2014; Konge et al. 2012; Beard 2005; Diwadkar et al. 2009; King et al. 2015; Thomsen et al. 2015; Preisler et al. 2015; Pedersen et al. 2014; Konge et al. 2013; Sedlack 2011; Sedlack & Coyle 2016; de Montbrun et al. 2015; Vassiliou et al. 2013; N. N. MD et al. 2015; Svendsen 2014). Participants are scored as pass or fail (competent or non-competent) by judges and the cut-off point is the intersection of these two groups based on a global rating scale or task-specific checklist scoring system. There has been an increased use of participant-centred methodology in more recent literature. This may be due to the fact that these methods allow expert judges to use global rating tools to determine pass/fail checklist scores. In procedural assessment, this is important as it allows experts to observe the subjects' performance and give an overall rating of pass/fail, or competent/non-competent, which can then be applied to a method of assessment, for example OSATS(Martin et al. 1997) or Global Operative Assessment of Laparoscopic Skills (GOALS)(Vassiliou et al. 2005). This allows for a standard that, although determined by expert opinion, still employs evidence-based scoring systems. Additionally, the literature implies that participant-centred methodology is applicable to 'real-life' procedural assessments, where conditions are not always controllable, unlike

simulation or OSCE environments(Beard 2005; Diwadkar et al. 2009; Vassiliou et al. 2013).

A similar participant-centred method described in the literature was used sparingly in the articles reviewed. The borderline group method and borderline regression method(Kissin et al. 2013; de Montbrun et al. 2016) employ an approach similar to the contrasting group method, the difference being that participants are assigned a pass, fail or borderline designation. The cut-off point is determined as the mean score of this borderline group(Kissin et al. 2013). This approach has limitations(Livingston & Zieky 1982). Participants may be labelled as borderline if a judge is unfamiliar with their procedural technique or if their judgement is based on something not measured by the test(Livingston & Zieky 1982). This may account for its limited use in medical/surgical procedural assessments.

The second most commonly used method of standard setting was the Angoff method, an item-centred technique(Cohen et al. 2013; Wayne et al. 2007; Huang et al. 2009; Wayne et al. 2008; Barsuk, Cohen, Vozenilek, et al. 2012; Barsuk, Cohen, Caprio, et al. 2012; Burch et al. 2005; Barsuk et al. 2009; Walzak et al. 2015; Yudkowsky et al. 2014; E. N. T. MD et al. 2014; Jelovsek et al. 2010; I. C. Green et al. 2013). This method was developed to establish cut-off points in assessment, and traditionally has been applied to written assessments(Norcini 2003). In the present review it was used mainly in simulation assessment and situations where steps of a task were predetermined and external factors controlled(Walzak et al. 2015; Wayne et al. 2007). Unlike participant-centred methods, this suggests that the use of item-centred standard setting is appropriate in these types of study, because, in order to predetermine the level of the borderline trainee, there must be complete standardization of the task. Therefore, item-centred methods may be best employed in simulation-based assessment where there is complete control over all non-surgeon factors.

There was great discrepancy among included studies in terms of participant experience level, study design and assessment purpose, making it difficult to pool study

data and perform an analysis of the combined studies. Although this made performing a meaningful meta-analysis impractical, it does lend credence to the applicability of this methodology across multiple arenas of assessment. Many studies used standard setting as a means of determining trainee progression, from the simulator to the bedside(Jacobsen et al. 2015; Yudkowsky et al. 2014; Preisler et al. 2015; Pedersen et al. 2014). It is crucial that trainees in procedural specialties meet a performance standard before being introduced to patient care, in order to maximize patient safety(Yudkowsky et al. 2014). The use of absolute standard-setting methods to define these benchmarks avoids the employment of an arbitrary norm-based method, and ensures content validity through the use of expert judges.

For an absolute standard to be credible, it must involve judges who are not only content experts in the field, but who also have undergone formal training in use of the assessment tool or method(Downing et al. 2010). The first step is selection of judges who are content experts in the field being assessed(Verheggen et al. 2008; Jaeger 1989). In the present review, this often meant experienced surgeons or those who performed the tested procedure frequently in high volume, but there was heterogeneity in how included studies defined the expertise of their judges. Judges need to be trained in the specific standard-setting method being used in the assessment(Cizek 1996). Although there is no single definition on how to train judges, the main steps include provision of written information on the exercise, discussion of the purpose of the standard-setting process, encouraging discussion of key concepts, and allowing practice opportunities with material similar to that in the actual test(Livingston & Zieky 1982; Cizek 1996). Subtle differences between experts in technique and ‘style’ may influence rating scores, and ratings from multiple judges control for the inherent subjectivity of a global rating system. Although some studies described their judges as being trained in standard setting, the nature and duration of this training was not specified. This detracted from the quality of this literature as the reproducibility is limited by the lack of clarity regarding judge selection and the process of judge training. The number of judges needed is dependent on the standard-setting method being employed(Jaeger 1989). For item-centred standards, between eight and ten judges are

usually required(Downing et al. 2010). For participant-centred standards, the accepted method is to have three to five judges watch each performance and deem it a pass or fail(Cizek 1996).

Of the three concepts described in Messick's conceptual framework(Messick 1994) that are used in the MERSQI(Reed et al. 2007), the majority of studies received full marks for content and relationship to other variables. Inter-rater reliability was not reported as consistently. This probably reflects the nature of standard-setting studies, which often use assessment metrics that have shown intraclass correlation (ICC) in previous studies. Failure to report the rater ICC is a possible source of confounding in these articles. Overall, transparent methodology used in most studies demonstrated that, when used correctly, standard-setting exercises were valid methods of setting defensible benchmarks in competency, performance and mastery.

There are some limitations to this review. The identified studies involved heterogeneous outcome measurements, which made pooling of the data in a meta-analysis, or comparing absolute standards with relative ones in a statistically meaningful way, impossible. At the same time, there was homogeneity in the methodologies of the articles included (as this article reviewed only two types of standard setting). This made it difficult, with quality assessment metrics such as MERSQI, to identify clearly which articles were of greater quality than others. Additionally, although many included studies used assessment instruments that had previously been used extensively and were likely to have adhered to many, if not all, domains of Messick's conceptual framework of validity, the authors of the articles did not address this specifically. This impacts on study quality, as seen in the overall MERSQI scores in *Table 4*. Although these studies show the applicability of absolute standard-setting methodologies in procedural assessment, in their current form these methods are not yet ready for implementation in high-stakes assessment. For educators and accreditors to be able to apply these standards reliably in summative assessments, there must be further explanations and standardization in the way judges are selected and trained in these methodologies.

10.7 Measuring Quality in Surgical Care

Understanding how best to capture and quantify surgical quality is a complex issue, one that is influenced by patient, physician, and healthcare system-level factors. While daunting, it is essential to adequately measure the quality of care delivered, in order to identify how best to improve the care provided to patients and optimize their experience in the healthcare system. A huge amount of money is spent annually by healthcare providers in Canada, the UK, and the US each year in order to collect, store, and abstract these data. Yet, it is apparent that despite these efforts we do not yet truly have an adequate understanding how to best provide hospitals, physicians, or patients with a meaningful representation of their performance, that will lead to continuous quality improvement and improved patient care.

Current methods of measuring the quality of care provided in surgery can be distilled into two primary categories: those that measure the processes in which care is delivered, or *process* measures, and the outcomes of the care provided by surgeons and the hospitals in which they work, or *outcome* measures. Process measures assess the manner in which a physician or hospital carry out clinical care, such as adherence to clinical guidelines or resource stewardship.

10.7.1 Benchmarking Quality of Care

Without clear standards or benchmarks, it is difficult to compare outcome and process measures across multiple healthcare providers. Statistical techniques must be applied in order for accurately comparisons of these metrics across heterogeneous patient populations, and these primarily focus on adjustments for patient ‘case-mix’ variation. In most cases, this involves the use of a multivariable regression model that includes the quality measures of interest alongside known patient confounders.

10.7.2 Currently used surrogates of surgeon quality

In lieu of collecting large amounts of process or outcomes data, researchers have sought to identify readily available surgeon factors that are closely associated with high quality care. Two types of such measures have been identified across multiple types of operations: annual procedural volume, and the surgeon's level of specialization in training.

The 'volume-outcome' relationship has been described at length in the literature, for many surgical procedures, both for cancerous and benign disease. This can be framed from an educational perspective as the 'learning curve' of a surgeon, or in a broader sense as the annual number of an operation-type completed by an individual surgeon. It is widely accepted in surgical education that while procedural volume may not be an ideal marker of competency, there is a trend toward improved technical skill with increased exposure and completion of a given operation. Procedural and anatomical knowledge, along with technical and non-technical skill acquisition likely drive this relationship in training. Volume has been shown to correlate with trainee competency in many surgical procedures. As discussed in previous subchapters, the initial approach to assessment of surgical trainees under the CBD curriculum relied on residents completing a minimum number of a given procedure, as a surrogate for competency. Similarly, the evidence supports the concept that when a practicing surgeon adopts a new operative approach or technique, there is an improvement in their surgical outcomes over an initial number of cases, which varies across procedure types. Different from the trainee's learning curve, this improvement in outcomes is less likely a product of improving procedural knowledge or non-technical skills, but more likely relates to improving familiarity with new equipment or technical skills, specific to the operative approach in question. This concept has been shown in RAS across multiple procedures, in particular RARP. Data supports the use of annual procedural volume as a surrogate for performance in practicing surgeons as well, likely related to the upkeep of technical skills, clinical decision making, and appropriate patient selection related to that given procedure.

Specialization has been explored as a surrogate for quality in a more limited manner in surgery. Evidence in orthopedic surgery (Norwood procedure), radical cystectomy, and radical prostatectomy all supports the concept that surgeons with subspecialty training in a given field of expertise are likely to have improved surgical outcomes. Obviously, this relationship likely is due to an increased exposure to more advanced or complex operations during this additional training period, and perhaps to an exposure to multiple operative approaches or techniques, outside of residency training.

10.7.3 Limitations of Currently Used Surrogates

While surrogate measures of quality are readily available and often simple to interpret, they likely are not adequate as highly accurate measures of surgical quality. In order to control for patient variables that may confound the predictive relationship between these surrogates and the outcome of interest, complex case-mix adjustment is undertaken. This manipulation of population-level data has been met with criticism by the academic community, as perhaps inadequate for capturing the heterogeneity of a given patient population. Although procedural volume and level of specialization have been shown to associate significantly with patient outcomes over large numbers of patients, the amount of variation explained by these factors may be insufficient for their use in reimbursement or credentialing practices.

10.7.4 The Skill-Outcome Relationship in Surgery

The volume or level of specialization of a surgeon ultimately serves as a proxy of their technical and non-technical skills, in the context of a specific procedure. It would seem practical to simply measure these factors directly, but as discussed above, this has historically been a challenging endeavour. However, growing evidence points to the ability of surgical skill assessments to associate significantly with patient outcomes

across a number of procedures. Birkmeyer and colleagues published a highly influential paper in 2013, that used peer-review of surgical video to stratify 20 participating surgeons into groups based on technical skill ratings using the OSATS scale. This study found significant differences in short-term postoperative complication rates between these groups of surgeons. This article was followed by studies from Hogg and Fecso that examined this skill-outcome relationship in pancreaticoduodenectomy and laparoscopic gastric cancer surgery, respectively.

This wave of high impact literature has spurred new interest in linking the skill of a surgeon or surgical team with patient safety or outcomes in a growing number of procedures, as highlighted in a recent systematic review from Fecso et al. However, few studies have tried to directly investigate the ability of a rating scale to independently predict the outcome of an individual surgical case, adjusting for patient and surgeon factors, and this remains a key line of investigation in surgical education and quality improvement.

10.7.4.1 Current Evidence in Radical Prostatectomy

Whether surgical performance can be used as a measure of quality in radical prostatectomy has been explored to varying degrees over nearly the past 20 years. Walsh published a noteworthy study in 2000 that identified four surgical techniques that led to improved sexual function at 18 months post-prostatectomy, using video collected intraoperatively to do so. Five years later, a French group used intraoperative video from laparoscopic radical prostatectomy (LRP) cases to study the impact of surgical technique on PSM, identifying and analyzing the technical factors that led to specific instances of PSM. Finally, Paterson et al found that in a prospective cohort of LRP cases, the skill of the surgeon as quantified by a novel scoring tool was an independent predictor of continence at three months.

10.8 Continuing Professional Development in Surgery

Despite minimal evidence examining the role of physician or surgeon competency on maintaining satisfactory patient care, continuing professional development (CPD) and continuing medical education (CME) programs have been in place internationally for over 40 years (Krause & Illich 1977). In principal, these assessment programs have been implemented as a means of ensuring that physicians and surgeons in practice continue to meet the minimum standard of their profession. However, with few exceptions, these activities have traditionally relied on subjective, self-assessments of one's own competency, without a formal examinations of technical or non-technical skills. Additionally, practices differ substantially both internationally and within Canada, leaving the potential for wide variations in surgeon competency across jurisdictions.

10.8.1 Current Recertification Practices

In Canada, recertification practices depend on where a practice is located. Nationally, the maintenance of certification (MOC) system employed by the RCPSC uses a five year cycle-based process that requires surgeons to present evidence of CME activity in the form of 'credits' acquired through participation at educational and academic conferences or courses. Some provincial colleges ask physicians to obtain documentation of self or peer-evaluation from colleagues based on the CanMEDS roles (Frank et al. 2015), and other provinces offer specific educational courses or examinations to complete their CME requirements. Completing these minimum standards of CME is not always straightforward. Urologists in Canada, especially those in a rural or community setting, have encountered barriers to participating in CME activities, citing issues with funding, coverage of clinical duties, and a lack of incentive or compensation for such requirements (Mahmood 2015).

CME and CPD activities vary considerably internationally. In the United Kingdom, all physicians must be ‘revalidated’ by the General Medical Council (GMC) every five years. This process involves a appointed senior doctor within a healthcare institution acting as the appraiser, formatively examining the professional development of the physician in question, in addition to a summative assessment of their progress over the five year period. The appraiser must account for any significant safety or professionalism events, the physician’s involvement in QI activities within their institution, and feedback from colleagues and patients, both positive and negative. In the United States, MOC is based on the six ACGME core competencies, and is evaluated across a four-part framework by the American Board of Medical Specialties (ABMS) (Specialties 2015). This process must be undertaken every ten years, and includes a formal written examination of the candidates medical knowledge related to their specialty, as well as participation in a minimum amount of CME per-cycle. In addition, the ABMS will account for any interval evidence of professional misconduct in their recertification decision.

11 Research Hypotheses and Study Aims

11.1 Thesis Purpose

This thesis will provide evidence toward outcome-based assessments and benchmarking in robotic surgical skill for the evaluation of surgical trainees in a competency-based framework and for incorporation into credentialing practices.

11.2 Hypotheses

1. Current methods of assessing technical skill in robotic surgery lack key sources of validity evidence, including the ability to predict or account for variations in patient outcomes.

2. Surgeon technical performance is a significant predictor of postoperative outcomes in robotic-assisted radical prostatectomy.

3. Standards can be set that reliably and accurately identify surgical technical performances that fail to provide patients with a reasonable probability of a satisfactory postoperative outcome.

11.3 Study Aims

Aim 1: To systematically review the currently used methods of assessing technical skill in urologic robotic surgery

i. Gather and categorize the technical skill assessment tools used in urologic robotic surgery, across all levels of training and experience;
ii. Examine the validity evidence supporting these assessment tools, using a structured and contemporary approach to test validity; and
iii. Identify the assessment tools with the most supporting evidence, for use in subsequent research.

Aim 2: To systematically review the currently used methods of setting performance standards in procedural technical skill

i. Establish the types of standard setting methods used in the medical education literature, including the environments where these assessments take place;
ii. Examine the manner in which assessment judges are selected, and trained for these standard setting exercises; and
iii. Examine the methodological rigour and overall quality of this literature.

Aim 3: To understand the relationship between surgical technical performance and postoperative outcomes in robotic-assisted radical prostatectomy

i. Evaluate the ability of an established metric of surgical technical skill in robotic surgery to predict clinically relevant patient outcomes, in a single-surgeon, retrospective cohort of robotic-assisted radical prostatectomy patients; and
ii. Further validate these findings in a prospective cohort of patients, using a multi-surgeon, multi-centre study design.

Aim 4: To create a novel approach to standard setting in technical performance, using the skill-outcome relationship as the benchmarking construct.

i. Use statistical modelling, rather than simple consensus, as the basis for creating standards in technical skill; and
ii. Describe this methodology using a staged approach, with a preliminary model created from retrospective data, followed by prospective validation.

12 Surgeon Performance Predicts Early Continence after Robotic-Assisted Radical Prostatectomy

12.1 Introduction

Until recently, the correlation between surgeon technical ability and postsurgical patient outcomes had not been scientifically proven. Birkmeyer and colleagues used objective ratings of surgeon skill using intraoperative laparoscopic video was found to be associated with early postoperative complications in bariatric surgery(Birkmeyer et al. 2013). Prior to this publication, efforts to improve patient outcomes in surgery took the form of system-based interventions, such as the surgical safety checklist(Haynes et al. 2009). At a time when surgeons find themselves under increased scrutiny from the public

and policy makers, these findings have spurred the academic community to further investigate the role of surgeon performance in determining patient outcomes. While many studies have opted to use readily available surrogates of surgeon performance, such as level of specialization(Bhindi et al. 2014) and procedural volume(Zevin et al. 2012; Trinh et al. 2013), only a small number have used direct observation of operative technical performance for this purpose. The link between surgeon ability and patient outcome has not been made regarding robotic surgery, an operative approach used in a growing number of surgical specialties.

The number of yearly radical prostatectomies performed with robotic-assistance has increased steadily since its introduction 16 years ago(Stitzenberg et al. 2012). While robotic-assisted radical prostatectomy (RARP) has not been shown to offer lower rates of positive surgical margins when compared to open radical prostatectomy (ORP), the literature suggests it imparts improved and earlier postoperative urinary and sexual function over traditional approaches(Hakimi et al. 2009; S. C. Kim et al. 2011). These *functional* outcomes are important in affecting the patient's quality of life (QoL) following surgery, in addition to placing a significant financial burden on the patient and healthcare system(Resnick et al. 2013). Urinary continence at three months postoperatively has commonly been cited as an appropriate mark for assessment of QoL after RARP (J. J. Kim et al. 2012; Abraham et al. 2010). Despite the precision offered by robotic surgery, there remains significant inter-hospital and inter-surgeon variation in RARP outcomes(Vickers et al. 2011). While there are known patient characteristics associated with incontinence and sexual dysfunction following radical prostatectomy, the role of the surgeon's technical performance in determining these outcomes has not been directly studied and remains unclear. This gap in knowledge is underscored by the absence of standardized training and accreditation in robotic surgery, in spite of its widespread adoption by urologists(J. Y. Lee et al. 2011).

Walsh and colleagues previously investigated the role of surgical technique in preserving postoperative sexual function in radical prostatectomy(Walsh et al. 2000). Through semi-structured analysis of intraoperative video footage, they identified four technical maneuvers that correlate with superior postoperative sexual function. With the

widespread adoption of minimally invasive surgery (MIS), the ability to capture video in the operating room has become much more feasible. A more recent study(Paterson et al. 2016) concluded that peer-review of video from extraperitoneal laparoscopic radical prostatectomy (ELRP) is weakly correlative with early postoperative continence, and this early finding supports our hypothesis.

We hypothesize that the technical performance of the operating surgeon is predictive of a patient's postoperative functional outcome in RARP. The 'return to continence' after RARP has been identified as a key contributor to the QoL of patients following surgery(Lavigueur-Blouin et al. 2015). We used blinded observation and objective evaluation of surgeon performance to understand whether technical performance is associated with urinary continence in patients three months after undergoing RARP.

12.2 Methods

Study Design and Subjects

We conducted a retrospective, one-to-one matched case-control study, examining the role of surgeon performance in predicting return to continence at three months postoperatively. Our study sample was drawn from a prospectively collected database of RARP patients at the University of British Columbia (CREB #H09-01628). At the time of this study, the surgeon had completed 512 RARP cases, and this database includes the entirety of the surgeon's RARP experience (case 1 to case 512). The database captures all clinically relevant preoperative information, including demographical data, clinical findings, biopsy results, and reports from any imaging studies. In addition, the database contains perioperative details, such as operative time, estimated blood loss, and intraoperative findings such as prostate median lobe, and extent of periprostatic nerve-sparing. Finally, the database contains postoperative details such as sexual function, oncological outcomes, additional or adjuvant treatments, and urinary function at 3, 6, and 12 months after RARP. Operative video for each case consisted of the intracoporeal

endoscopic view only, containing no patient or surgeon identifiable information, and no audio. Ethics approval for this study was granted by the institutional clinical research ethics board (CREB #H16-00940).

Patients identified as being incontinent at three months postoperatively were randomly selected from the database and matched with continent patients in age, body mass index (BMI), surgical reconstruction of anterior or posterior fascia, preoperative International Prostate Symptom Score (IPSS), and prostate weight (drawn from the pathology report). In addition, cases were matched by surgeon case number, to account for position on the learning curve. All putative controls were identified from the database, then randomly selected to limit selection bias. Continence was defined as not requiring an underwear pad, or the use of a single daily precautionary pad only. No exclusion criteria were applied; all patients who underwent RARP, with or without pelvic lymphadenectomy, were included. A single surgeon operated on all patients in our sample (SLG). Trainees acted as the bedside assistant, or occasionally completed simple steps of the operation as primary surgeon (i.e. bladder drop). All included patients had a urethral catheter placed postoperatively, removed 7 or more days after surgery.

Video Analysis

A single rater, with expertise in the procedure content, as well as orientation and training in the assessment methods, performed the all the video ratings. Cases were de-identified and labeled with study codes, meaning the rater was blinded to all patient characteristics and outcomes. The operative video was evaluated using two separate rating scales: the Global Evaluative Assessment of Robotic Skill (GEARS, appendix-2)(Goh et al. 2012), and the Generic Error Rating Tool (GERT, appendix-3)(Bonrath, Zevin, et al. 2013). GEARS consisted of five, 5-point Likert scale domains of robotic technical skill: depth perception, bimanual dexterity, efficiency, force sensitivity, and robotic control (the ‘autonomy’ domain of the GEARS score was not rated, as it could not be assessed from the video footage)(Ghani et al. 2016). A higher GEARS score is associated with superior surgical skill. GERT records the number of errors committed by the primary surgeon and bedside assistant throughout the procedure, and during all

operative steps. Errors are defined as any deviation from the expected course of the operation, as defined by Bonrath et al(Bonrath, Zevin, et al. 2013). Intraoperative events, including bleeding, broken sutures, and inadvertent tearing of tissue, were also captured. Both of these assessment instruments have evidence to support their use in assessing surgeon performance in the operating room, as outlined by Messick's Contemporary Framework(Messick 1975). The cases were broken down into steps defined by the Pasadena Consensus(Montorsi et al. 2012). GEARS scores were calculated for six main operative steps: Bladder Drop, Endopelvic Fascia, Bladder Neck Dissection, Seminal Vesicle Dissection, Pedicle Division and Preservation of Neurovascular Bundles, and Urethrovesical Anastamosis. The mean score of these steps was used as the total GEARS score for the case.

Statistical Analysis

Frequency statistics were calculated for patient demographical data and surgeon scores, and a Shapiro-Wilk test with $p > .05$ was used to define normal distribution. Univariate analysis was conducted to test for statistically significant differences in assessment scores between incontinent and continent cohorts across all variables, using Independent Sample T-Tests and Mann-Whitney U testing as appropriate. Spearman Rho tests were conducted to asses for correlation between errors, events, and GEARS scores. A variable-selection strategy was used to create multivariate models. Binary logistic regression was conducted to calculate odds ratios (OR) and 95% confidence intervals for significant predictors on univariate analysis, and clinically relevant co-variates. Statistical significance was set at $p < .05$ based on a two-tailed comparison. Statistical analyses were performed using SPSS Statistics version 23 (IBM, NY, USA).

12.3 Results

Patient Sample

24 patients deemed to be incontinent at 3 months postoperatively were randomly selected from the database, and were matched for age, BMI, fascial reconstruction, IPSS, prostate weight, and case number with 23 continent patients (Table-10). Additional patient demographic and perioperative data are listed in Table-10, including estimated blood loss (EBL) and length of stay (LOS). The cases were drawn equally from across the surgeon's learning curve (continent vs. incontinent mean case number 271.13 vs. 288.54, p = .60).

Table-10: Patient Demographics

	Continent (n=23)		Incontinent (n=24)		
	Mean, Median or Frequency (SD or IQR, %)				p-value
BMI	26.44	(2.36)	26.64	(2.83)	.64
Age	60.13	(7.82)	63.91	(5.60)	.06
Smoking					
Yes	4	(17.3)	2	(8.3)	.39
Previous	5	(21.7)	3	(12.5)	
No	14	(60.8)	19	(79.1)	
ASA Score					
0	3	(13.0)	2	(8.3)	.68
1	5	(21.7)	9	(37.5)	
2	12	(52.1)	10	(41.6)	
3	3	(13.0)	3	(12.5)	
Previous TURP					
Yes	1	(4.3)	2	(8.3)	.58
No	22	(95.6)	22	(91.6)	
IPSS	9.13	(9.09)	11.20	(8.54)	.16
SHIM	22.70	(7.12)	18.08	(8.96)	.11
Case Number	271.13	(169.60)	288.54	(163.17)	.60
Pre-operative PSA	5.21	(2.32)	8.10	(8.92)	.16
Prostate Weight	50.28	(19.03)	56.16	(18.48)	.90
Median Lobe					
Yes	6	(26.1)	7	(29.1)	.58
No	17	(73.9)	16	(70.9)	
EBL (ml)	331.73	(368.97)	195.83	(151.02)	.40
Foley Days	10.59	(7.43)	8.87	(2.58)	.76
Length of Stay	1.70	(1.46)	1.26	(0.75)	.11
Anterior Hitch					
Yes	15	(65.2)	13	(54.1)	.44
No	8	(34.7)	11	(45.8)	
Rocco Suture					.45
Yes	9	(39.1)	12	(50.0)	
No	14	(60.8)	12	(50.0)	
Anastomotic Leak	3	(13.0)	0	(0.0)	.07
Bladder Neck Contracture	0	(0.0)	2	(8.3)	.16
Positive Surgical Margin					
Negative	18	(78.3)	19	(79.2)	.99
<2mm	2	(8.7)	2	(8.3)	
>2mm	3	(13.0)	3	(12.5)	
Mean Nerve Spare (%)	68	(31)	57	(24)	.12

SD = Standard deviation, IQR = Interquartile Range, BMI = Body mass index, EBL = Estimated blood loss, PSA = Prostate Specific Antigen, IPSS = International Prostate Symptom Score, SHIM = Sexual Health Inventory for Men, ASA = American Society of Anesthesia, TURP = Transurethral resection of prostate

Correlations

There was a strong inverse correlation between mean overall GEARS and total GERT (r = -.68, p < .001), meaning an increased number of errors per case was associated with lower GEARS scores. Previous studies using GERT also showed similar correlation with other global rating scales.(Bonrath, Zevin, et al. 2013) Total adverse events showed moderate inverse correlation with GEARS scores (r = -.42, p < .05), and strong positive correlation with GERT (r = .76, p < .001)

GERT

No difference in total errors (p = .97) or events (p =.88) was seen between the study cohorts (Table-11). During the bladder neck dissection step, there were a greater number of total errors committed in the incontinent group (Mdn 6.1 vs. 3.3), but this difference did not reach statistical significance (p = .07).

GEARS

Mean GEARS score was higher in the continent group (Mdn 19.6 v. 19.0, p = .007, Table-11). When broken down by operative step, statistically significant differences in GEARS scores are seen during the bladder neck dissection (Mdn 20.0 vs. 19.0, p = .01) and urethrovesical anastomosis (Mdn 20.0 vs. 18.6, p = .02) steps.

Table-11: Differences in GEARS and GERT scores between continent and incontinent cohorts

Assessment Tool	Step	Continent (n=23) Mean or Median (SD or IQR)	Incontinent (n=24) Mean or Median (SD or IQR)	(p-value)
GEARS (Score /25)	Bladder Drop	19.0 (2.3)	20.0 (3.0)	.78
	Endopelvic Dissection	20.0 (3.3)	20.0 (1.5)	.12
	Bladder Neck	**20.0 (3.0)**	**19 (2.2)**	**.01**
	Seminal Vesicles	20.0 (3.3)	18.0 (2.0)	.07
	Pedicles & Neurovascular Bundle	19.0 (2.3)	19.0 (2.0)	.40
	Urethrovesical Anastamosis	**20.0 (2.3)**	**18.6 (1.2)**	**.02**
	Overall	**19.6 (1.7)**	**19.0 (1.5)**	**.02**
GERT (Number of Errors)	Bladder Drop	5.8 (3.5)	4.1 (2.6)	.20
	Endopelvic Dissection	2.7 (1.7)	2.5 (1.1)	.53
	Bladder Neck	3.3 (3.8)	6.1 (5.5)	.07
	Seminal Vesicles	8.7 (5.3)	5.9 (4.4)	.20
	Pedicles & Neurovascular Bundle	8.5 (5.5)	8.0 (5.0)	.64
	Urethrovesical Anastamosis	5.7 (6.4)	3.7 (2.5)	.61
	Overall	41.0 (17.0)	35.0 (11.0)	.97

Bold values indicate statistical significance ($p<0.05$)
IQR = Interquartile Range
GEARS = *Global Evaluative Assessment of Robotic Skill*
GERT = *Generic Error Rating Tool*
RACE = *Robotic Anastomosis Competency Evaluation*

Multivariate Analysis

Binary logistic regression models were constructed to understand whether mean GEARS score was predictive of continence at three-months postoperatively, controlling for clinically and statistically relevant confounders (Table-12). Patient age (p = .06), BMI (p = .79), and prostate weight (p = .29) were included as co-variates, based on their status in the literature as patient predictors of continence. Total GEARS score was independently predictive of continence (*OR* = .55, 95% CI .33-.91). A sub-step analysis was performed, and both bladder neck dissection (*OR* = .69, 95% CI .51-.94) and urethrovesical anastomosis (*OR* = .70, 95% CI .50-.97) GEARS scores were independently predictive as well.

Table-12: Binary Logistic Regression Models

	OR	95% CI		p-value
Total GEARS	.55	.33	.91	p=.02
Age	1.09	.98	1.23	
BMI	.98	.77	1.26	
Prostate Weight	1.00	.97	1.04	
Bladder Neck GEARS	.69	.51	.94	p=.01
Age	1.11	.99	1.24	
BMI	.98	.79	1.31	
Prostate Weight	1.00	.97	1.04	
Urethrovesical Anastomosis GEARS	.70	.51	.94	p=.03
Age	1.12	.97	1.20	
BMI	.98	.78	1.04	
Prostate Weight	1.00	.97	1.04	

GEARS = Global Evaluative Assessment of Robotic Skill
OR = Odds Ratio, CI = Confidence Interval, BMI = Body mass index

12.4 Discussion

This study is the first to demonstrate a predictive relationship between surgeon technical performance and patient outcomes in RARP. We found that assessment of surgical video by a trained analyst using a valid metric of surgical performance can independently predict the early return to continence at three-months for patients undergoing RARP. On sub-step analysis, GEARS scores at two key steps of the procedure, bladder neck dissection and urethrovesical anastomosis, were also independently associated with continence. Our study's findings may have significant implications for training and assessment, as well as for stakeholders in the accreditation and education of robotic surgeons. These findings, along with similar studies in bariatric(Birkmeyer et al. 2013) and pancreatic(Hogg et al. 2016) surgery, continue to clarify the important role that a surgeon's technical performance plays in shaping significant postoperative outcomes.

GERT was not associated with continence in this study, despite the expected finding of a strong negative correlation with total GEARS score. No differences in total numbers of errors or events were seen at any sub-step of the procedure as well. This may be due to insufficient precision of the tool. Many errors and events captured by the GERT may be clinically insignificant, as the instrument was designed to limit rater subjectivity and therefore captures a broad definition of error, including 'near-misses'(Bonrath, Zevin, et al. 2013; Bonrath, Gordon, et al. 2015). Similarly, intraoperative events that are recorded with the tool may not impact patient safety, but instead reflect the skill of the surgeon, as the strong correlation with GEARS would suggest. Further development of this instrument is necessary prior to its use in this context moving forward, including a method of classifying errors and events by severity, rather than just a combined total.

The objective analysis of surgeon technical performance shows promise as a metric of surgical quality(Dimick & Varban 2015). With the introduction of quality-based payment (QBP) in the United States there is a pressing need for robust measures of

performance, at both the hospital and physician levels(Dimick & Greenberg 2013). Currently, policy makers rely on the retrospective analysis of population level data in order to assess quality across healthcare providers, using often complex methods of statistical risk and case-mix adjustment to account for differences in predictors of patient adverse outcomes(Maas et al. 2013). However, the inherent bias of these types of analyses have caused some to question the adequacy of these methods for high-stake decision making such as public reporting and physician compensation(Ban et al. 2016). As patients demand to be privy to surgeon outcomes data, the onus is on providers to ensure that the metrics we provide the public and payers are both accurate and reliable. Our evidence presented here demonstrates an alternative method of measuring quality that can be performed using existing instruments. Importantly, while previous evidence has shown that the GEARS assessment can differentiate between surgeons of different levels of training(Aghazadeh et al. 2015), this study demonstrates its ability to distinguish differences in the performance of an individual surgeon. This adds to the validity 'argument' for the use of such assessments for surgeons on an iterative basis to track progression along the learning curve(Cook et al. 2015).

The findings of this study have implications for trainees and educators, in addition to surgeons wishing to improve the outcomes for their patients. Walsh indicated that as urologists we can improve our surgical performance through self-assessment(Walsh et al. 2000), and these findings indicate that this exercise will benefit our patients as well. Peer-assessment of technical ability is a growing field, and the application of structured coaching models adopted from other industries form the basis for multiple recent studies(Min et al. 2015; Bonrath, Dedy, et al. 2015). Trainees and surgeons of all levels of experience can benefit from coaching, and this study gives educators the specific assessment methodologies and operative steps to implement to influence RARP outcomes. In addition, with the arrival of competency-based medical education (CBME) in surgical training, the evidence presented here could prove invaluable to educators in designing formative and summative assessments for making judgments of competency in robotic surgery(Szasz et al. 2014). Finally, it is important to note that this study adds to the validity evidence for the GEARS tool. Surgeon *skill* is unlikely to change from case-

to-case, but these findings would suggest that *performance* does, and this objective assessment tool is able to differentiate between the operative performances of a single surgeon.

There are limitations to our study. The single-surgeon, patient-matched design limits confounding, but these findings must be confirmed through a prospective study, including multiple surgeons and if possible, multiple institutions. Secondly, we used a single rater for our video-analysis. However, the current consensus is that multiple raters are not necessary when using a tool that has been repeatedly validated in the literature(Ghani et al. 2016; Aghazadeh et al. 2015), as reliability is a property of the assessment tool, not the assessment itself. Training and orienting the analyst to the assessment methods enhances rating accuracy. Thirdly, while the GERT has been used to assess laparoscopic procedures in other surgical specialties, its application to urology and robotics lacks validity evidence. This may explain its inability to distinguish between clinical outcomes. Finally, while our findings indicate that surgeon performance accounts for some of the variability in patient outcomes, it is likely that there are currently unquantifiable intraoperative factors that make a case more surgically challenging. These technical and non-technical variables must be further elucidated and included in future comparable analyses.

12.5 Conclusion

This study has provided the first evidence that surgeon performance in RARP is associated with early return-to-continence. Assessment of technical performance by a trained analyst using a valid metric of surgeon performance may be an independent predictor of patient outcomes in RARP. Error rating was not predictive, but the application of a severity scale may improve its predictive properties. This hypothesis-generating retrospective study requires prospective validation using multiple surgeons and institutions. Stakeholders in the accreditation and training of minimally invasive urologists

should heed this study's findings when designing summative assessments and educational interventions in RARP.

13 A Novel Method of Standard Setting Using Patient Outcomes

13.1 Introduction

In the new paradigm of Competency-Based Medical Education (CBME), the evaluation of surgical competency is moving from the classroom to the operating room. In an outcome-based curriculum, the formative assessment of surgical competencies through simulation or workplace-based assessments (SBA or WBA) must be paired with summative evaluations that ultimately determine whether an individual has reached the threshold to proceed in their training.(Frank et al. 2010) This concept is echoed later in the career of a surgeon, in the form of accreditation and continuing medical education (CME).(Pradarelli et al. 2015) As we discover the unique relationship between surgeon performance in the operating room and clinically significant patient outcomes, there is a growing demand for the assessment of these parameters in surgeons already in independent practice.(Dimick & Greenberg 2013) Ultimately, the responsibility to set defensible standards in technical and non-technical skill falls to surgeon educators and regulating bodies. We owe it to our patients and the public to ensure that we not only accredit trainees who perform above a predetermined benchmark, but that we strive to monitor the performance of surgeons throughout their careers.

Traditionally, methodologies that determine the pass mark in an assessment are categorized as creating either absolute or relative standards.(Norcini 2003) Relative standards are based on the performance of an index group, typically experts in the area being assessed. After quantifying this index group's performance, using a chosen assessment method, the standard is set in relation to their scores, for example using standard deviations or percentiles. In contrast, absolute standards are the result of an expert consensus process, either through conceptualizing how a 'borderline' participant

would score on a given station ('item-centred'), or by making a judgement on the observed performance of each assessment participant ('participant-centred').(Downing et al. 2006) Absolute standards are often seen as being more suitable for high stakes assessment, as they allow educators to apply an 'external' set of criteria to an assessment, that reflect the purpose of the assessment. For example, in the case of CBME these criteria would be those that define the performance of a 'competent' trainee.

Setting standards in surgical competency has been examined previously in the literature. In a recently published systematic review of absolute standard setting for the evaluation of procedural skill,(M. G. Goldenberg, Garbens, et al. 2017) our group found that although using educational methods to set competency benchmarks in assessments is feasible, the existing evidence in this field is of moderate quality, and is lacking in validity evidence. We found that the literature has explored these methodologies primarily in the simulation environment, where real-world consequences are not assessable. There are exceptions to this, including Beard et al(Beard 2005) using the Contrasting Groups Method to distinguish between experts and non-experts in a cohort of vascular surgeons of different training levels. Similarly, Szasz and colleagues(Szasz et al. 2016) applied this same standard setting methodology to a prospective cohort of trainees performing laparoscopic cholecystectomy, demonstrating that their absolute standard was closely correlated to other variables that predict trainee competency, such as level of training and surgical volume. In contrast to traditional absolute and relative standard setting methods, the Receiver Operating Characteristic (ROC) curve analysis has been used only sparingly as a statistical means of setting standards in procedural competency. In a landmark study carried out by Fraser et al,(Fraser et al. 2003) an ROC curve was created to establish a pass/fail score for their MISTELS training system, determining the sensitivity and specificity of various cutoff points in their assessment tool for simulated laparoscopy.

Although established methodology has been successfully adopted from traditional assessment formats to evaluate trainee procedural competence, these

methods rely on expert consensus or external criteria to define the assessment construct. To improve the credibility of benchmarks for high-stakes assessments of surgical performance, we describe a data-driven approach to setting standards in the operating room, through a combination of predictive statistical modelling and ROC curve analysis.

13.2 Methods

The present paper proposes a novel multi-factorial model to establish outcome-based procedural standards that account for both the variation in procedural WBAs in the operating room. This method can be applied to any cohort study of a procedure with a clinically relevant, quantifiable outcome of interest, with independent variables that include a measure(s) or physician performance and/or patient confounders. To illustrate this method, we used data derived from our recently published pilot study that demonstrated the predictive relationship between surgeon technical performance and post-operative continence in robotic-assisted radical prostatectomy (RARP).(M. G. Goldenberg, L. Goldenberg, et al. 2017)

Model Data

The model was applied to a retrospective, matched cohort analysis of patients at a single institution who underwent RARP by a single surgeon between 2008-2015.(M. G. Goldenberg, L. Goldenberg, et al. 2017) Consented patients had their clinically relevant information previously abstracted from the medical record and recorded in a de-identified, prospective database, including unedited intraoperative, intracorporeal video of the procedure. An expert rater analyzed footage and use validated metrics of performance (Global Evaluative Assessment of Robotic Skills (GEARS) and the Generic Error Rating Tool (GERT)) to evaluate surgeon performance, blinded to the outcome of each case. Similar to other studies evaluating video taken from operating room, the GEARS rubric in this analysis omitted the ‘autonomy’ domain, making 25 points the maximum score possible.(Ghani et al. 2016) The primary outcome of the analysis was

continence at 3 months post-operatively, which has been described previously as an important indicator of functional surgical outcome and quality of life (QoL).

Statistical Analysis and Benchmarking

Frequency-distribution analysis was undertaken for the outcome variables of interest, and any patient characteristics that are known to predict for, or are associated with, the primary outcomes. Appropriate univariate testing (i.e. simple regression, t-test) was used to detect differences in ratings and confounding patient characteristics. A *variable-selection strategy* was used to create a multivariate model that included the three highest cited patient factors that contribute to continence. Multivariate ROC analysis was performed by first building a multivariate logistic regression model, controlling for patient characteristics. The regression model was then used to determine predicted probabilities, which was used as the test variable in the ROC analysis against the clinical outcome (state variable). The optimal cut-off score that best predicts three-month continence was then determined from the ROC table. To accomplish this, we used the optimal value of predicted probability, termed Youden's index, which maximizes the tradeoff between sensitivity and specificity values. Statistical Analyses were performed using SPSS version 24 (NY, USA).

13.3 Results

The analysis yielded a binary regression model that included: patient age, body mass index (BMI), prostate volume, and surgeon technical performance as measured by the mean GEARS score for the procedure, calculated as the mean score for the 6 steps analyzed (Table-1). In this model, performance was found to be independently predictive of continence at three months post-operatively (OR 0.55, p = .02). The ROC curve analysis yielded an area under the curve (AUC) of 0.75 (Figure-1). Youden's Index was found to be at a predicted probability of 0.35, which corresponds to a sensitivity of 0.96 and a specificity of 0.52 (Table-2). We selected these parameters for the purposes of demonstrating this methodology, but alternative probabilities can be

chosen with sensitivity and specificity that better reflect the underlying assessment construct.

Table-13: Results from the multivariable regression analysis used in the pilot study

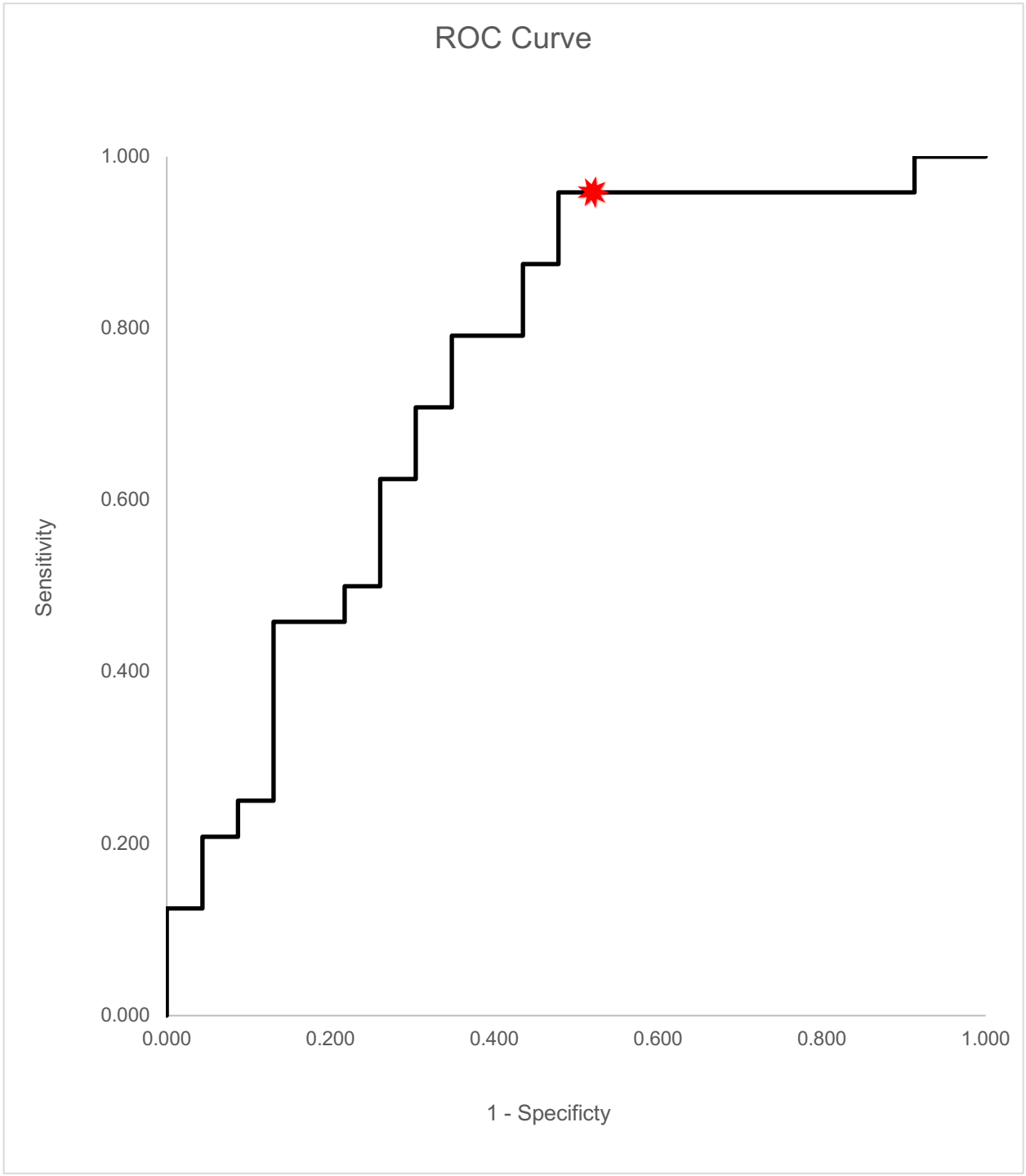
	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Total GEARS	-.606	.261	5.396	1	.020	.546	.327	.910
Age	.094	.057	2.773	1	.096	1.099	.983	1.227
BMI	-.019	.126	.023	1	.879	.981	.766	1.256
Prostate Weight	.006	.018	.106	1	.745	1.006	.971	1.042
Constant	6.257	6.641	.888	1	.346	521.741		

Table-14: Truncated receiver operating characteristic (ROC) curve co-ordinates and their corresponding predictive capabilities

	Positive if Greater Than or Equal To	Sensitivity	Specificity
Total GEARS Predicted Probability	0.0000000	1.000	0.000
	.0473255	1.000	.043
	↑	↑	↑
	.2771744	.958	.391
	.2816372	.958	.435
	.3104328	.958	.478
	.3504087	**.958**	**.522**
	.3847808	.917	.522
	.4298695	.875	.522
	.4574342	.875	.565
	↓	↓	↓
	.8351015	.042	1.000
	1.0000000	0.000	1.000

Youden's Index (optimal trade-off between sensitivity and specificity) is bolded in the table

Figure-3: The ROC curve, using the model's predicted probability as the test variable and the clinical outcomes of interest as the state variable. The red star indicates Youden's Index in this example, the cutoff probability that maximizes sensitivity and specificity in the model.



Using this chosen predicted probability of 0.35, we next reverse engineered the regression formula to solve for the GEARS score, rather than the outcome (Figure-2). In doing so, we could determine the required GEARS score that corresponds to the chosen predicted probability of continence at three months post-operatively. Thus, by inputting the case-specific patient parameters in the regression formula (age, BMI, prostate volume), we can determine the benchmark, that if achieved, will provide that specific patient with a 65% probability of being continent following RARP. As shown in Table-3, a 62 year old patient, with a BMI of 28 and a prostate weight of 150 grams would require a GEARS score of 21.53 out of a possible 25 points. As the GEARS scale only provides whole numbers, this would be rounded up or down depending on the nature of the assessment.

Figure-4: 'Reverse engineered' regression formula, to calculate GEARS score required to give a 35% probability of an adverse outcome

$$\text{GEARS} = \frac{6.257 + 0.094(\text{Age}) - 0.019(\text{BMI}) + 0.006(\text{Prostate Weight}) - \ln[0.35-(1-0.35)]}{0.606}$$

Table-15: The rearranged regression equation allows for patient characteristics to adjust the performance score benchmark, based on an assessment-specific chosen predicted probability

Patient Age	62
Body Mass Index (BMI)	28
Prostate Weight	150
Constant	1
Probability of Incontinence at 3 months	35.0%
Total GEARS Score Needed	**21.53**

In order to provide a 62-year old patient with a BMI of 28, a prostate size of 150g with a 65% chance of being continent at 3 months postoperatively., a total GEARS score of 21.53 is required

13.4 Discussion

This is the first description of a method of standard setting that is based on a trainee or physician performance as a predictor of clinically important patient outcomes. The adaptation of regression modelling and ROC curves to determine a context and patient-specific performance standard is unique to medical education.

Like most statistical analyses, this methodology is predicated on certain assumptions that must be satisfied prior to the standard being calculated. First, the assessment tool used must have sufficient validity evidence for the evaluation's purpose and context. Educators can use a framework, such as Messick's or Kane's,(Cook & Beckman 2006) to compose a 'validity argument' in support of the assessment rubric selected. In the case of procedural competency, a task-specific checklist (TSC), or as in our example a global rating scale (GRS), can be selected that addresses the underlying construct of the assessment. A component of contemporary validity arguments is whether the outcome of the assessment (i.e. pass versus fail) has any impact on the trainee, clinical care, or the wider health care system (referred to as 'extrapolation'/'consequences'). This principle has been underexplored in healthcare assessments, and represents the second assumption of the present methodology.(Hatala et al. 2015) To set a benchmark that is outcome-based, the relationship between the assessment tool and the outcome of interest must be delineated in a statistical analysis or predictive model. Emerging evidence supports this concept across multiple surgical subspecialties, including bariatric surgery,(Birkmeyer et al. 2013) general surgery(Hogg et al. 2016) and as described in this manuscript, urology.(M. G. Goldenberg, L. Goldenberg, et al. 2017) All these studies use datasets and analyses that could be applied to set performance standards in their individual surgical procedures, using well-established GRS such as the Objective Structured Assessment of Surgical Skills (OSATS)(Martin et al. 1997) and the GEARS rubric.(Goh et al. 2012)

The strength of this standard setting methodology is its applicability to WBAs. Unlike SBAs, evaluations of competency in the workplace must be responsive to the heterogeneity that is inherent in patient care.(Norcini 2005) The ability to standardize the assessment environment, a strength of simulation and virtual reality platforms, is not achievable in WBAs, making it essential that the 'pass-mark' be context-specific. By inputting patient parameters into the model, the benchmark is set to reflect the specific patient in front of the trainee or surgeon being evaluated. In essence, this standard setting method allows for accurate comparisons to be made between WBAs, irrespective of changes in patient factors from one assessment environment to the next. Similarly, it is important that trainees and clinicians be held to a standard that demonstrates their ability to 'elevate their game' when the clinical context demands it. As WBAs become ingrained in CBME curricula and culture, their role will vary depending on the assessment purpose. Emphasis on sensitivity versus specificity may differ from summative to formative assessments, and our method allows educators to weigh these two factors appropriately, using the ROC analysis. In this way, the benchmark that results from this methodology is both patient-specific, context-specific, and assessment type-specific.

As competency-based assessments become mandatory in surgical training programs, resident program directors and surgeon educators will be tasked with completing regular formal evaluations of trainee operative performance. If surgical skill is to be integrated into summative assessments for progression through a training program or for accreditation, we feel the standards or benchmarks set in these assessments should be evidence-based and clinically relevant. As the predictive relationship between surgeon performance and significant patient outcomes continues to be better understood in multiple procedures, this methodology will have growing applicability. For example, this methodology could be applied to Birkmeyer et al's finding that technical performance in laparoscopic gastric bypass is predictive of postoperative complications.(Birkmeyer et al. 2013) Choosing this endpoint, a multivariate model using OSATS as the measure of surgical performance could be create that includes patient factor covariates. The same process described in this article could then be

applied and a patient-specific standard could be set. This benchmark could be used as a pass-mark for senior trainees or fellows seeking accreditation in bariatric surgery, or for surgeons wishing to renew their licensure in this procedure, thereby helping to standardize patient outcomes or identify those surgeons who may require remediation.

Beyond its utility in postgraduate training and CBME, the use of predictive models to set technical performance standards would have implications for continuing medical education (CME) practices. Currently, technical skills are not formally assessed at the time of surgeon credentialing,(Tam et al. 2017) and evidence suggests that may have negative repercussions for patient safety.(Pradarelli et al. 2015) This issue is particularly important when new surgical technologies or techniques are introduced into the surgical community, and lessons from the rapid, unchecked dissemination of robotic-assisted surgery highlight this problem.(Zorn et al. 2009; Y. L. Lee et al. 2011) This novel assessment methodology presents one way to address this matter, by holding surgeons to a standard the directly relates to measures of patient safety. Further investigation of the role of technical and non-technical performance in optimizing surgical safety will allow for the development of more comprehensive, statistically-derived benchmarks.

A recognized obstacle for the successful implementation of novel assessment methods is a lack of physician engagement and 'buy-in.' As surgeons, we are evidence-driven in our clinical decision making and patient care. As such, it is often difficult to accept standards or accreditation practices that are the result of consensus alone. A unique principle behind this method is that it is the result from predictive modelling, and hard clinical endpoints that are relevant to clinicians. We hypothesize that this characteristic will increase the perceived credibility of performance standards set using this methodology.

There are limitations that must be noted in the design and use of this standard setting method. First, the predictive model used to determine the standard may not address all the variance in the outcome of interest. The present model uses a GRS of

technical skill as a measure of surgeon performance, but it is likely that non-technical skills and other unmeasured confounders account for a proportion of the variation in predicting continence following RARP. However, as this field evolves, predictive models will likely include more of these metrics, and our ability to quantify many intraoperative factors, such as decision-making and judgment, will improve. A second limitation relates to the procedure-specific nature of the standard that is set. Educators wishing to utilize this methodology in the assessment of a given procedure must first undertake the work to design a predictive model. As more procedures are being investigated in this way, emerging data will be generated to apply the present method in a growing number of clinical specialties. It should be noted that the data used in the example described here comes from a single surgeon study. Finally, it must be acknowledged that in our provided example, the chosen predicted probability of 35% corresponds to a false-positive rate of 48%, and this may not be acceptable based on the outcome chosen and the nature of the assessment. This limitation is specific to the multivariate model from our pilot study used as an example, and ongoing work includes validating this model with a prospective, multi-surgeon study. Future applications of this method will yield different ROC analyses, and different trade-offs in sensitivity and specificity to be considered.

13.5 Conclusion

The present paper describes a novel method of standard setting, that utilizes predictive models to statistically determine the context, patient, and assessment-specific benchmark for a given procedural task. The applicability of this method is growing, as more literature is produced supporting the underlying assumptions of the method, in particular the ability of the intraoperative surgeon performance to predict clinically important outcomes. This method can be used across the spectrum of CBME for both formative and summative assessments, and the data-driven nature of the approach may increase its credibility amongst trainees, surgeons, and educators.

14 Surgeon intraoperative performance predicts patient outcomes in robotic-assisted radical prostatectomy: a prospective, multicenter analysis

14.1 Introduction

A growing body of evidence supports the notion that variations in postoperative patient outcomes is in part due to differences in technical skills between surgeons.(Fecso et al. 2016) Although other measures of surgeon technical performance have been described in the literature such as surgical volume(Trinh et al. 2013) and degree of specialization,(Bhindi et al. 2014) these ultimately represent surrogates for the ability of the surgeon to safely and successfully complete a given procedure. Analysis of large, population-level datasets has been successful in identifying clinically important quality indicators across surgical specialties,(Maggard-Gibbons 2014; Lawson et al. 2017) but these efforts lack the granularity to provide surgical teams with the data they need to improve the quality of care they deliver. Recent efforts reported in the literature have used direct observation of procedures, in both real-time and through video-capture, to quantify the capacity of a surgeon or the entire surgical team to provide high quality intraoperative care.(Fecso et al. 2017; Hogg et al. 2016; Birkmeyer et al. 2013; M. G. Goldenberg, L. Goldenberg, et al. 2017) With increased attention being paid to discrepant outcomes for patients across geographic, political, and societal boundaries, the ability accurately compare surgeon skill and technique has been identified as a way of broadly standardizing care for patients undergoing potentially life-altering procedures.(Tam et al. 2017)

Robotic-assisted surgery (RAS) has been singled out in recent times as an example of unregulated dissemination of surgical technology, to the potential detriment of patient safety.(J. Y. Lee et al. 2011) Industry-driven growth of this platform across the Western world has shaped an uneven landscape of care delivery for patients undergoing RAS procedures.(Lloyd 2011) Robotic-assisted radical prostatectomy (RARP) has

expanded over more than 15 years to be the gold-standard approach to surgical treatment of localized prostate cancer.(Lowrance & Parekh 2012) Outcomes following RARP have implications beyond cancer control, with short and long-term urinary and sexual dysfunction directly impacting the quality of life in these patients.(Alemozaffar et al. 2015) Despite the high-stakes nature of this procedure, the absence of standardized accreditation practices has potentially contributed to increased medicolegal concern around robotic-surgery, particularly in lower-volume or non-academic hospitals.(Zorn et al. 2009) Video assessment of robotic technical skill has been proposed in recent years as a potential solution to these issues, as a means of both supplementing the accreditation and proctoring process for new robotic surgeons, as well as enhancing training during residency and fellowships.(O'Mahoney et al. 2016; M. G. Goldenberg, Jung, et al. 2017)

When assessing surgical performance, it is essential to use instruments with validity evidence supporting their use in a given context. [18,19] However, we still do not fully understand how the scores from currently used robotic surgical assessment tools relate to patient outcomes, and this gap in evidence limits the ability of these tools to be used to make informed decisions regarding the ability of trainees and surgeons to safely and competently perform RAS.(M. G. Goldenberg et al. 2018)

To better recognize the utility of performance-based assessment in RAS, we must first understand the predictive properties of the available methodologies. In this study, we aimed to validate our previous retrospective findings, which demonstrated a significant association between surgeon performance scores and clinically-relevant patient outcomes, using a multicenter, multi-surgeon, prospective study design. Confirmation of this predictive relationship between skill and outcome would further add to the validity evidence supporting video-based assessments of technical skill as a potential intervention to limit the variability of postoperative outcomes across the spectrum of robotic surgical care.

14.2 Methods

Study Setting

This is a prospective, multi-surgeon study conducted at three teaching hospitals with established robotic surgery programmes. RARP procedures analyzed in this study were captured over a 9-month period (March 2016 – November 2016), with a follow-up period of one year to allow outcomes to mature. This study was approved by the research and ethics boards at all participating hospitals, with data sharing agreements in place to allow transfer and centralization of clinical data.

Study Participants

Patients undergoing RARP at the three participating hospitals over the study period were eligible for enrollment. Patients were included after obtaining informed consent. No exclusion criteria were enforced, and any patient undergoing RARP with or without lymph node dissection was eligible for inclusion in the study.

Surgeons conducting RARP cases at the three participating hospitals were consented at the start of the study period. All included staff surgeons completed a minimum of 5 RARP cases annually. In addition, surgical trainees (residents and fellows) were included in the study, with no restrictions on their experience or annual exposure to RARP. Demographic data from all faculty and trainee surgeons were collected at the outset of the study. Data collected included previous robotic experience, level of training, and RARP exposure.

Data Collection

All surgeon and patient participants were de-identified and randomly-generated study codes were assigned to each participant following the consent process. Intraoperative video recordings were obtained directly from the daVinci Vision Tower™,

and did not contain images or information that could identify the patient or participating surgeons. Audio was not captured during the case.

Baseline patient clinical information including age, body mass index (BMI), prostate specific antigen (PSA), prostate biopsy data (gland volume, Gleason grade), and clinical tumour stage, were collected at the time of surgery. Patients agreeing to participate completed the urinary and sexual domains of the Expanded Prostate Cancer Index Composite-26 (EPIC-26) questionnaire(Szymanski et al. 2010) prior to surgery, to establish baseline functional status. Intraoperative data including estimated blood loss (EBL), presence of a median lobe, degree of nerve-sparing, dorsal venous complex (DVC) management, bladder neck reconstruction, posterior approach, and total operating time were collected. Postoperative outcomes were collected retrospectively at one year following the surgery, through chart review. The pathology report and medical record were reviewed at this time to identify final grade and stage and functional outcomes (stress incontinence and erectile dysfunction).

Video Analysis

Each un-edited video contained the entire procedure from insertion of the trocars to insertion of the Jackson-Pratt drain. To objectively assess the performance of the surgical team, two global rating scores (GRS) were used, the Global Evaluative Assessment of Robotic Skills (GEARS),(Goh et al. 2012) and the Prostatectomy Assessment Competency Evaluation (PACE).(Hussein et al. 2016) These instruments have been described in the literature previously, with validity evidence supporting their use in the reliable and accurate assessment of intraoperative robotic skill.(M. G. Goldenberg, L. Goldenberg, et al. 2017; Hussein et al. 2016) GEARS scores were assigned to 6 operative steps: bladder drop and prostate preparation, bladder neck dissection, seminal vesicle dissection (SV), pedicle and neurovascular bundle dissection (NVB), apical dissection, and urethrovesical anastomosis (UVA). Lymph node dissection was not scored in the analysis. As previously described,(M. G. Goldenberg, L. Goldenberg, et al. 2017; Ghani et al. 2016) modified version of the GEARS was used to

rate each operative step, and the mean of these step scores was used as the total case GEARS score. The GEARS score consists of five, 5-point Likert scale domains of technical skill, including depth perception, bimanual dexterity, efficiency, force sensitivity, and robotic control. The PACE score is also Likert-scale based, but instead assesses how well the surgeon carries out 10 domains of skill over seven operative steps. A mean score for each case was calculated using the scores from each domain assessed in the PACE.

Three independent, blinded analysts participated (authors M.G., A.G., and H.S.). All analysts underwent a structured, standardized orientation with the assessment tools, the purpose of the study, and the rating process. Interrater reliability (IRR) statistics were calculated for both GEARS and PACE scores, with Cronbach's Alpha calculated on cases analyzed by at least two analysts. Scores from each GRS Likert-scale domain were used in the analysis.

Outcomes

Three pre-defined patient outcomes were chosen *a priori* based on their clinical significance and impact on patient quality of life.(Cooperberg et al. 2012) Positive surgical margin (PSM) was defined as any tumor extending to the cut tissue plane, and was not restricted by size or location. Two functional outcomes were chosen. Urinary continence at 3 months postoperatively, defined as completely dry or the use of a single 'safety' pad, and erectile function at 12 months postoperatively, defined as erections adequate for sexual activity (including masturbation) with or without the use of pharmacotherapy (i.e. phosphodiesterase inhibitors). Urinary and sexual function were chosen as primary outcomes given their known contribution to a patient's quality of life following radical prostatectomy (Wei et al. 2000).

Statistical Analysis

Distributions of surgeon and patient demographic data, as well as performance scores were explored using histograms and tested for normality using Shapiro-Wilk tests

(p > .05 defined normal distribution). Surgeon performance scores were compared between case experience and training levels using Kruskal-Wallis tests. Preoperative and intraoperative variables were compared in a bivariate analysis, and performance scores were compared across the three outcomes of interest using Chi-Square for categorical variables and Mann-Whitney U for continuous variables. A similar bivariate analysis using Mann-Whitney U tests compared performance scores between patients with and without PSM, continence at 3 months, and erectile function at 12 months. Patients were clustered around these endpoints to limit confounding, with patients with pre-surgical erectile dysfunction or urinary incontinence removed in these respective analyses. Binary logistic regression was used to test the significance of performance score predictors in multivariable models for each of the outcomes. Patient factor-covariates in these models were selected based on their known association with the outcome of interest. In a sensitivity analysis, fixed effects were accounted for by including hospital volume and surgeon experience variables into the multivariable model. Finally, cross-validation of the models was performed using a traditional K-fold validation method. Statistical significance was set at 0.05 based on a two-tailed comparison. Statistical analyses were performed using SPSS Statistics v24 (IBM, NY), and R Statistics (Vienna, AUT).

14.3 Results

Study Participants

A total of 31 surgeons participated in the study, including 11 (35.5%) staff surgeons, 14 (45.2%) fellows, and 6 (19.4%) residents. Each staff surgeon contributed between 1 and 31 cases to the study over the 9 month collection period (Table-16). A broad range of surgeon experience was seen amongst trainee and surgeon participants. Staff surgeons ranged in pre-study exposure to RARP, with 4 (4.3%) of cases completed by surgeons with less than 30 cases experience, and 44 (47.8%) of cases completed by surgeons with greater than 250 completed cases prior to the study period. Amongst trainees, 13 had exposure to less than 10 RARP cases at the outset of the study, and only 2 had participated in more than 30 cases. Prior experience in open prostatectomy

ranged from less than 30 cases (7, 22.6%) to more than 250 cases (10, 32.3%) completed, amongst all participants including trainees.

Table-16: Surgeon and Trainee Demographics

	N	%
Participating Surgeons		
Staff	11	35.5
Fellow	14	45.2
Resident	6	19.4
Surgeon RARP Experience		
<30 cases	4	4.3
30-100	19	20.7
101-250	25	27.2
>250	44	47.8
Surgeon Open Prostatectomy Experience		
<30 cases	0	0
30-100	0	0
101-250	22	23.9
>250	70	76.1
Hospital Volume		
<100 cases annually	9	9.8
≥100 cases annually	83	90.2
Primary Surgeon		
1	12	13.0
2	31	33.8
3	4	4.3
4	2	2.2
5	13	14.1
6	1	1.1
7	10	9.8
8	7	7.6
9	4	4.3
10	9	9.8
Trainee (Resident and Fellow) RARP Experience		
<10 cases	13	68.4
10-30 cases	4	21.1
>30 cases	2	10.5

93 patients were provided consent and participated in the study, with no patients approached refusing to participate. Of these, one patient's video file was damaged, leaving 92 patients in the final analysis (Table-17). Median operating time was 180 minutes (IQR 152.5-229.5), with a median estimated blood loss of 250 millilitres (IQR 200.5-337.5). The cohort consisted primarily of intermediate CAPRA risk patients (72, 78.2%), with 72 (78.3%) Gleason grade 7 and only one (1.1%) clinical T3 tumor in the

cohort. Median preoperative PSA was 9.0 (IQR 4.8-10.5). At baseline, 7 (7.6%) patients reported urinary incontinence, defined as leaking of urine requiring more than a single security pad daily, and 15 (16.3%) reported erectile dysfunction prior to surgery. Postoperatively, there was a single case requiring blood transfusion, and 10 (10.9%) patients presented to the emergency department or were admitted to hospital within 30 days of surgery. Median length of stay was 1 day (IQR 1-2), with only 9 (9.7%) cases requiring prolonged hospital admission (3 or more days). Median time with urinary catheter postoperatively was 13 days (IQR 12-15), and 7 (8.0%) patients had a UVA leak diagnosed with either pelvic drain fluid creatinine level or cystogram.

Table-17: Preoperative Demographics of Patients Included in the Study

	Mean/Median	SD/IQR
Age	61.2 mean	(6.93) SD
Blood Loss (ml)	250	200-350
Days to Catheter Removal (84)	13.5	3
Total OR Time	167	144-209
PPB	0.369/36.9%	0.25-0.58/ 25%-58%

N=92	No	%
Prior Surgery		
No	72	79.1
Yes	18	19.8
Missing	1	1.1
Median Lobe		
No	73	79.3
Yes	18	19.6
Missing	1	1.1
Nerve Spare		
No	15	16.3
Yes	76	82.6
Missing	1	1.1
DVC Method		
Open	58	63.0
Suture	32	34.8
Missing	2	2.2
Posterior Reconstruction		
No	53	57.6
Yes	39	42.4

Length of Stay		
1	47	51.1
2	36	39.1
≥3	9	9.8
Urethrovesical Anastomotic Leak		
No	81	88.0
Yes	7	7.6
Missing	4	4.3
Post-Operative Blood Transfusion		
No	87	94.6
Yes	1	1.1
Missing	4	4.3
Readmission		
No	78	84.8
Yes	10	10.9
Missing	4	4.3
Prostate Specific Antigen		
<4.4	20	21.7
4.5-8.9	40	43.5
≥9	32	33.7
Prostate Volume		
≤46.5	45	48.9
>46.5	47	51.1
BMI		
≤27.5	49	53.3
>27.5	42	45.6
Missing	1	1.1
Preoperative Stress Incontinence		
No	86	93.5
Yes	6	6.5
Preoperative Erectile Dysfunction		
No	77	83.7
Yes	15	16.3

At 3 months postoperatively, 54 (58.7%) patients reported no stress urinary incontinence, with no pad or the use of a single security pad daily (Table-19). At one year, 42 (45.7%) patients described erections sufficient for intercourse or masturbation, with the use of no medications or phosphodiesterase inhibitors only. On final pathology, 32 (34.8%) of patients had T3 disease, with 15 (48.3%) PSM in this subgroup. 60 (65.2%) patients had T2 disease at final pathology, and 13 (21.7%) of these patients were found to have a PSM on pathology review.

Table-19: Postoperative Patient Outcomes

	No	%
Stress Incontinence at 3 Months		
No (Dry or Safety Pad)	54	58.7
Yes	38	41.3
Erectile Dysfunction at 12 Months		
No	42	45.7
Yes	50	54.3
Staging (After Biopsy)		
T2a	13	14.1
T2b	12	13.1
T2c	66	71.7
T3	1	1.1
Biopsy Gleason Score		
6	8	8.7
7	72	78.3
≥8	12	13.0
Pathological Gleason Grade		
6	9	9.8
7	72	78.2
≥8	11	12.0
Pathological Stage		
≤T2b	19	20.6
T2c	41	44.6
≥T3	31	33.7
Missing	1	1.1
Surgical Margins		
pT2 (n=60)		
Negative	47	78.3
Positive	13	21.7
T3 (n=31)		
Negative	16	51.7
Positive	15	48.3
Overall		
Negative	64	69.6
Positive	28	30.4

Surgical Performance Ratings

Twenty cases (21.7%) were rated by at least two analysts, in order to calculate IRR statistics for performance ratings. GEARS scores demonstrated *good* IRR, with a Cronbach's Alpha of 0.726, PACE scores showed excellent IRR, with a Cronbach's Alpha of 0.877.

Comparing step scores across training levels, GEARS scores at the bladder drop ($p = .01$), bladder neck ($p = .03$), and seminal vesicles (SV) ($p = .01$) steps were all higher in staff surgeons compared to fellows and residents. PACE scores differed significantly across training levels as well, at the UVA step (needle entry, $p = .01$, needle driving & tissue trauma, $p = .01$, and urethrovesical approximation, $p = .04$). In surgeons with greater than 250 cases completed at the outset of the study, GEARS scores were higher at the bladder drop ($p = .01$), bladder neck ($p = .02$), and NVB ($p < .01$), and PACE scores were higher at the bladder drop ($p = .01$), prostate preparation ($p = .03$), UVA (needle entry, $p = .04$, needle driving & tissue trauma, $p = .05$) steps in these surgeons, when compared to those with fewer than 250 cases.

Bivariate Analysis

Clinically relevant patient characteristics (Table-19), and performance scores (Table-20), were compared across dichotomous primary outcome variables. Prostate volume was significantly larger in patients with incontinence at 3 months postoperatively ($p = .03$), nerve-sparing was used less in patients with erectile dysfunction at 12 months postoperatively ($p = .02$), and tumor stage was significantly higher in patients with PSM ($p = .03$). GEARS scores were higher in continent patients overall ($p = .03$), and at the bladder neck ($p = .03$), NVB ($p = .03$), apical ($p = .03$), and UVA ($p < .01$) steps. Higher PACE scores in the continent cohort were found at the bladder neck ($p = .04$) and UVA (needle entry, $p < .01$, needle driving, $p = .03$) steps, as well overall ($p < .01$). Overall PACE scores were higher in patients with adequate erectile function one year postoperatively only ($p = .03$), with no single step significant on bivariate testing. Overall

and step GEARS scores were not significantly different between patients with and without erectile dysfunction at one year. Finally, patients without PSM had higher GEARS scores during the bladder drop (p = .03) step only, whereas PACE scores were higher in these patients at the SV (seminal vesicles, p = .03, posterior plane, p = .02) and apex (p = .02) steps, as well as overall (p = .02).

Table-19: Bivariate Analyses of Patient Factors by Subgroup

Continence (n=85)

	Continent at 3 Months (n=53)		Incontinent at 3 Months (n=32)		
	Mean	*SD*	*Mean*	*SD*	*p-value*
Age	61.8	8.4	60.6	5.9	.42
	n	*%*	*n*	*%*	*p-value*
BMI					
≤27.5	30	56.6	17	53.1	.46
>27.5	23	43.4	15	46.9	
Prostate Volume					
≤ 46.5	31	58.5	11	34.4	.03
> 46.5	22	41.5	22	65.6	
Nerve-Sparing					
No	7	13.2	6	19.4	.33
Yes	46	86.8	25	80.6	
DVC management					
Suture	31	58.5	23	71.9	.16
No Suture	22	41.5	9	28.1	
Median Lobe					
No	42	79.2	26	81.3	.38
Yes	11	20.8	5	15.7	
Posterior Reconstruction					
No	26	49.1	22	68.7	.06
Yes	27	50.9	10	31.3	

Note: p-values are from independent samples t-test (Age) and chi-square tests for all other categorical variables.

Erectile Function (n=77)

	Erectile Function at 12 Months (n=36)		Erectile Dysfunction at 12 Months (n=41)		
	Mean	*SD*	*Mean*	*SD*	*p-value*
Age	58.3	6.4	62.1	7.1	< .01
	n	*%*	*n*	*%*	*p-value*
BMI					
≤27.5	21	58.3	23	56.1	.52
>27.5	15	41.7	18	43.9	
Prostate Volume					
≤ 46.5	20	55.6	18	43.9	.21
> 46.5	16	44.4	23	56.1	
Nerve-Sparing					
No	2	5.6	10	24.4	.02
Yes	34	94.4	31	75.6	
Tumor Stage					
≤T2b	6	16.7	10	25.0	
T2c	20	55.6	16	40.0	.39
≥T3a	10	27.8	14	35.0	
Gleason Grade					
6	5	13.9	3	7.5	
7	29	80.6	31	77.5	.31
≥8	2	5.6	6	15.0	

Note: p-values are from independent samples t-test (Age) and chi-square tests for all other categorical variables.

Positive Surgical Margin (n=92)

	Negative Margins (n=64)		Positive Margins (n=28)		
	n	%	n	%	*p-value*
BMI					
≤27.5	32	50.8	17	60.7	.26
>27.5	31	49.2	11	39.3	
Prostate Volume					
≤ 46.5	32	50.0	13	46.4	.47
> 46.5	32	50.0	15	53.6	
Nerve-Sparing					
No	12	18.8	4	14.3	.42
Yes	52	81.3	24	85.7	
Tumor Stage					
≤T2b	16	25.4	3	10.7	.03
T2c	31	49.2	10	35.7	
≥T3a	16	25.4	15	53.6	
Gleason Grade					
6	6	9.5	3	10.7	.89
7	50	79.4	21	75.0	
≥8	7	11.1	4	14.3	
Prostate Specific Antigen					
<4.4	17	27.0	3	10.7	.19
4.5-8.9	27	42.9	13	46.4	
≥9	19	30.2	12	42.9	

Note: p-values are from chi-square tests

Table-20: Bivariate Analysis of Surgeon Performance

	Continence at 3 Months (n=85)					Potency at 12 Months (n=77)					Surgical Margin (n=92)				
	No (n=32)		Yes (n=52)			No (n=41)		Yes (n=36)			No (n=64)		Yes (n=28)		
	Median	IQR	Median	IQR	p-value	Median	IQR	Median	IQR	p-value	Median	IQR	Median	IQR	p-value
Bladder Drop															
GEARS	4.4	3.4-4.6	4.0	3.4-4.6	.33	4.1	3.4-4.6	4.0	3.6-4.8	.54	**4.4**	**3.4-4.8**	**3.8**	**3.4-4.5**	**.03**
PACE															
Bladder Drop	4.0	4.0-5.0	5.0	4.0-5.0	.57	4.0	4.0-5.0	5.0	4.0-5.0	.43	5.0	4.0-5.0	4.0	4.0-5.0	.24
Preparation of Prostate	4.0	3.0-5.0	5.0	4.0-5.0	.48	4.0	3.8-5.0	5.0	4.0-5.0	.27	5.0	4.0-5.0	4.0	3.0-5.0	.06
Bladder Neck															
GEARS	**4.0**	**3.4-5.0**	**4.6**	**4.2-4.8**	**.03**	4.4	3.8-4.8	4.6	4.2-4.8	.27	4.6	3.8-5	4.4	4.0-4.8	.43
PACE	**4.0**	**3.0-5.0**	**5.0**	**4.0-5.0**	**.04**	4.0	3.0-5.0	5.0	4.0-5.0	.28	4.0	4.0-5.0	4.0	3.5-5.0	.73
Seminal Vesicles															
GEARS	4.2	3.8-4.8	4.4	4.0-4.6	.39	4.3	4.0-4.6	4.4	3.8-4.8	.52	4.4	4.0-4.6	4.2	3.8-4.6	.24
PACE															
Seminal Vesicles	4.0	3.0-4.0	4.0	3.0-4.0	.68	4.0	3.0-4.0	4.0	3.0-4.0	.52	**4.0**	**3.0-4.0**	**3.0**	**3.0-4.0**	**.03**
Posterior Plane	4.0	3.0-5.0	4.0	3.0-5.0	.72	4.0	3.0-5.0	4.0	3.0-5.0	.81	**4.0**	**3.0-5.0**	**4.0**	**3.0-4.0**	**.02**
Neurovascular Bundle Dissection															
GEARS	**4.0**	**3.8-4.6**	**4.4**	**4.0-4.6**	**.03**	4.2	3.8-4.6	4.4	4.2-4.8	.16	4.4	3.8-4.6	4.2	3.8-4.6	.55
PACE	3.0	2.0-4.0	3.0	3.0-4.0	.08	3.0	2.0-4.0	4.0	3.0-4.0	.05	3.0	2.0-4.0	3.0	2.0-4.0	.59
Apical Dissection															
GEARS	**4.0**	**3.8-4.8**	**4.6**	**4.2-5.0**	**.03**	4.3	3.8-4.6	4.6	4.0-5.0	.11	4.4	3.8-5.0	4.4	3.8-4.7	.29
PACE	**4.0**	**3.0-5.0**	**4.0**	**4.0-5.0**	**.06**	4.0	3.0-5.0	5.0	4.0-5.0	.11	**4.0**	**4.0-5.0**	**4.0**	**3.0-4.0**	**.02**
Urethrovesical Anastomosis															
GEARS	**3.8**	**3.4-4.0**	**4.2**	**3.6-4.6**	**<.01**	3.8	3.6-4.5	4.0	3.4-4.4	.93	4.0	3.6-4.6	3.8	3.6-4.3	.46
PACE															
Needle Entry	**4.0**	**4.0-4.0**	**5.0**	**4.0-5.0**	**<.01**	4.0	4.0-5.0	5.0	4.0-5.0	.32	4.0	4.0-5.0	4.0	4.0-5.0	.33
Needle Driving	**4.0**	**3.0-4.0**	**4.0**	**4.0-5.0**	**.03**	4.0	3.0-5.0	4.0	4.0-5.0	.45	4.0	3.0-5.0	4.0	3.0-4.0	.14
Approximation	5.0	4.0-5.0	5.0	4.0-5.0	.65	5.0	4.0-5.0	5.0	4.0-5.0	.50	5.0	4.0-5.0	5.0	4.0-5.0	.92
Case Total															
GEARS	**4.1**	**3.9-4.4**	**4.3**	**4.1-4.5**	**.03**	4.2	3.9-4.3	4.3	4.1-4.5	.16	4.3	4.0-4.5	4.1	3.9-4.4	.09
PACE	**3.9**	**3.5-4.3**	**4.3**	**4.0-4.5**	**<.01**	**4.0**	**3.7-4.3**	**4.3**	**4.0-4.5**	**.03**	**4.3**	**3.9-4.5**	**4.0**	**3.6-4.3**	**.02**

Note: p-values are from Mann-Whitney tests

Multivariable Analysis

Binary logistic regression models were used to test the association between performance measures and each patient outcome, after adjusting for clinically relevant patient factors (Table-21). Only when the overall GEARS or PACE score was significantly different between groups on bivariate analysis was a multivariable analysis performed. After controlling for patient age, nerve-sparing status, prostate volume, patient BMI, and the use of posterior reconstruction, overall GEARS score (*OR* 3.5, *95% CI* 1.0 – 12.4, $p < .05$) and overall PACE score (*OR* 6.8, *95% CI* 1.8 – 24.7, $p < .01$) were independently predictive of postoperative continence, with areas under the curve (AUC) of 0.692 and 0.742 respectively. Overall PACE score was predictive of erectile function at 12 months (*OR* 5.6, *95% CI* 1.4 – 23.0, $p = .02$) after adjusting for patient age, nerve-sparing status, prostate volume, and BMI, with an AUC of 0.776. Finally, overall PACE score was an independent predictor of positive surgical margins (*OR* .27, *95% CI* .08 - .95, $p = .04$) after controlling for pathological stage, Gleason score, and preoperative PSA, with an AUC of 0.725.

Table-21: Binary Logistic Regression Models

Continence

	Odds Ratio	Confidence Interval	p-value
Overall GEARS	3.53	1.01-12.43	< .05
Age	0.98	0.92-1.05	.55
Nerve-Spare	0.89	0.21-3.75	.87
Volume	0.44	0.17-1.15	.09
BMI	0.83	0.31-2.24	.71
Posterior Reconstruction	1.82	0.66-5.03	.25
Overall PACE	6.80	1.87-24.68	< .01
Age	0.98	0.90-1.05	.53
Nerve-Spare	0.99	0.21-4.59	.99
Volume	0.40	0.15-1.09	.07
BMI	0.86	0.31-2.42	.78
Posterior Reconstruction	1.53	0.53-4.44	.43

Erectile Function

	Odds Ratio	Confidence Interval	p-value
Overall PACE	5.73	1.37-23.02	.02
Age	0.92	0.85-0.99	.04
Nerve-Spare	0.43	0.21-1.68	.11
Volume	5.05	0.82-32.97	.09
BMI	0.92	0.31-2.53	.88

Positive Surgical Margin

	Odds Ratio	Confidence Interval	p-value
Overall PACE	0.28	0.08-0.96	.02
≤ pT2b	Ref	Ref	.03
pT2c	2.52	0.55-11.60	.24
≥ pT3a	7.05	1.46-34.13	.02
Gleason 6	Ref	Ref	.89
Gleason 7	1.35	0.26-7.12	.73
≥ Gleason 8	1.00	0.13-7.67	.99
PSA < 4.5	Ref	Ref	.58
PSA 4.5-8.9	2.21	0.49-9.90	.30
PSA ≥9	1.81	0.36-9.22	.47

Sensitivity Analysis

To control for fixed hospital and surgeon effects, the experience of the primary surgeon and the annual hospital volume were included in the model. These variables were categorically coded, with 4 levels of surgeon experience (< 30, 30-100, 100-250, > 250), and 2 levels of annual hospital volume (≤ 100, >100). Overall PACE score remained a significant predictor in the continence (*OR* 10.8, *95% CI* 2.6 – 44.9, $p < .01$) and PSM (*OR* .17, *95% CI* .04 - .74, $p = .02$) cohorts, but not in the erectile function model (*OR* 3.5, *95% CI* .67 – 17.6, $p = .13$). Including hospital and surgeon factors in the model made overall GEARS score no longer a significant predictor (*OR* 3.7, *95% CI* .98 – 14.1, $p > .05$).

Model Validation

Cross-validation was performed to further test the predictive models. Overall PACE models in the continence and PSM cohorts underwent a traditional K-fold validation. Using 10 folds, the AUC of the continence model only decreased slightly from 0.742 to 0.740. However, the PSM model's AUC decreased more substantially in cross-validation, from 0.725 to 0.521.

14.4 Discussion

The present study supports our previous findings that surgical technical performance contributes significantly to variations in oncological and functional outcomes following RARP. (M. G. Goldenberg, L. Goldenberg, et al. 2017) Using a prospective multicenter study design with over 30 surgeons and trainees, we added important validity evidence for the use of the GEARS and PACE assessment scores as metrics of surgical quality in this setting.

These findings add to the growing body of literature that has linked technical skills with patient outcomes across many surgical specialties.(Fecso et al. 2016) Seminal work from Birkmeyer and colleagues(Birkmeyer et al. 2013) established this line of investigation as a potential means of bridging the quality gap in surgical care, but the use of video-review in radical prostatectomy goes back almost 20 years. Walsh et al.

described the use of video-capture in open surgery to recognize key technical manoeuvres that contribute to long term sexual function. (Walsh et al. 2000) In 2005, a French study used a similar study design in laparoscopic prostatectomy patients to identify the underlying segments of the dissection that led to a PSM. (Touijer et al. 2005) While informative and novel, the generalizability of these group's findings are limited by their retrospective and case-control study design. The present work provides evidence supporting the skill-outcome relationship in prostatectomy, by analyzing a multi-surgeon, prospective cohort of patients using strict educational and statistical methodologies. The significant relationships between metrics of technical performance and three distinct and clinically important outcomes strongly supports the use of these assessment strategies as not only educational tools, but potentially quality indicators for surgeons in practice.

These data provide a compelling argument for the use of the GEARS and PACE instruments in high-stakes assessment going forward. A competency-based surgical curriculum is already being used in multiple countries around the world, despite a lack of evidence that these outcome-based assessment strategies are clinically impactful.(Szasz et al. 2014) The present study demonstrates that in RARP, and potentially other RAS procedures, existing measures of technical skill can be effectively used to distinguish between surgical performances that will positively or negatively impact clinical outcomes. These findings should compel educators to implement evaluations of technical skill at all levels of in-training assessment, both for clinically meaningful feedback along the early learning curve and for evidence-based summative decision-making purposes.

In addition to educational practices, this study adds to a growing body of evidence supporting the addition of video-based technical skill assessments for surgeons in practice.(Tam et al. 2017) At a time of increasing scrutiny regarding the quality of healthcare delivery, the impetus is on the medical community to take a methodical, evidence-based approach to closing this quality gap, in part by minimizing the variation in outcomes across providers. (Lindenauer et al. 2014) Understanding the importance of surgical performance, beyond an anecdotal basis, must be a central tenant of credentialing practices. (Sachdeva & Russell 2007) Rather than leave important decisions

around the safe adoption of surgical procedures and proctoring practices to industry, clinicians and policy-makers in surgical care should drive the systems that ensure surgeons are safe to operate independently without compromising patient outcomes.(Pradarelli et al. 2015) This study allows stakeholders to take the next step in implementing standardized methods of credentialing surgeons, that uses objective measurement and granular evaluation of surgeon performance, for the benefit of patient outcomes and quality of life.

There are limitations to this work that warrant mention. Despite efforts to minimize type 2 error and maximize sample size, wide confidence intervals were seen in the multivariable analysis. This likely relates to our decision to look at the case-by-case association of technical performance and outcome, as opposed to applying a measure of *skill* (from a single or limited sample of video) to larger dataset, as done in similar previous work.(Birkmeyer et al. 2013) This study design allowed us to account for fluctuations in the *performance* of a surgeons across multiple procedural steps or cases, limiting the selection bias of using a single, surgeon-selected step or case. Additionally, it is imperative that we mention the unmeasured role of non-technical skills in this study.(Gostlow et al. 2017) Important factors such as teamwork, communication, and decision-making certainly play a role, both as a predictor of outcome and a confounder of technical performance.(Brunckhorst et al. 2015) The quantification of these important skills will continue to grow in the literature, and will certainly be studied in RAS and RARP. (Raison, Wood, et al. 2017) Finally, it is important to comment on the effect of the sensitivity analysis and cross-validation. It is likely that although these performance measures have a meaningful association with the outcomes collected in this study, the role of other patient, surgeon, and hospital-level factors must be acknowledged as equally impactful. Certain outcomes in this series, in particular erectile function, are simply difficult to predict based on the data collected, and this led to lower AUC values that decreased further in cross-validation.

14.5 Conclusion

This study provides key evidence supporting the use of technical skill assessments for both training and credentialing surgeons undertaking RARP procedures. These results confirm that performance in the operating room has a direct and likely independent predictive relationship with important patient outcomes, that may have implications for quality of life following RARP. While it is true that not all surgeons are necessarily able to reach the same levels of technical skill, this data supports the creation of quality benchmarks that define a minimum level of competency in surgical performance. Broad implementation of these assessment strategies may lead to a reduction in the variation of outcomes seen across the spectrum of RAS providers.

15 Evidence-based benchmarking in surgical performance: leveraging the skill-outcome relationship in procedural assessment

15.1 Introduction

In a competency-based training model, it is imperative to use assessments that are outcome-based, ensuring that a trainee or physician has demonstrated the necessary abilities of a proficient healthcare provider, and have displayed the capacity to carry out the core tasks of their area of practice.(Holmboe et al. 2010) Unfortunately, as we continue to build toward a foundation of competency-based medical education (CBME) in surgical training and beyond, we do not yet truly understand the implications that these assessments and their scores have on the patient care.(Hatala et al. 2015) As many of the studies supporting the benefits CBME have been completed in the theoretical space, the knowledge gap between assessment scores and clinical outcomes remains to be bridged.

A proposed model for facilitating the implementation of outcome-based assessments in CBME comes in the form of Entrustable Professional Assessments (EPA).(Cate et al. 2016) In surgical education, this has been proposed as a means of ensuring trainees' ability to safely carry out the essential surgical procedures of a given specialty, using a stepwise approach to skills assessment.(Pugh et al. 2017) This EPA framework relies on multiple assessments over the course of a trainee's residency, with multiple 'milestone' assessments that incorporate all the component skills needed to carry out a given procedure or clinical task. The question however remains as to the optimal method of determining when and how these milestone assessments should be carried out, and importantly what criteria should be used to identify which trainees can progress and which require further remediation.

The current literature that proposes methods of implementing an EPA framework in surgical skills assessment have primarily focused on consensus-based, educational outcomes (i.e. criteria that if met mean a trainee is 'competent' at a given task), and as of yet do not account for the safety or well-being of the patient being cared for.(Szasz et al. 2014) Methods used to benchmark scores in an assessment are adopted primarily from wider educational theory, and rely on human judgment to determine what criteria must be met to 'pass' an evaluation.(Downing et al. 2006) As these methods were not developed with medical education in mind, there is a disconnect between the benchmark being set and the clinical implications of allowing those trainees who meet this standard to independently carry out these procedures, often without monitoring for the entirety of their career.

Illuminating the relationship between surgical performance, particularly technical skills, and clinically important patient outcomes, remains a topic under investigation in many surgical specialties.(Fecso et al. 2016) Growing evidence supports the use of objective scoring tools to quantify surgical performance, and the ability of these scores to predict a proportion of the variation in certain outcomes across multiple procedure types.(Hogg et al. 2016; Birkmeyer et al. 2013; Fecso, Bhatti, et al. 2018; M. G. Goldenberg, L. Goldenberg, et al. 2017) This relationship has important implications for

surgical education as a whole, but particularly for assessments of technical skill in the operating room, during training and accreditation.

In this study, we further develop our previously published standard setting methodology, that leverages the association between assessment scores and patient outcomes to set benchmarks for educational use.(M. G. Goldenberg & Grantcharov 2017) We aim to further modify our statistical technique, through the use of procedural step-weighting and composite performance scores that improve the predictive properties of these metrics, and allow educators to better appreciate the importance of individual operative steps on clinically relevant outcomes.

15.2 Methods

Study Design

This study uses data from a prospective, multicenter cohort of patients undergoing RARP, from March-November 2016 in Toronto, Canada. Intracorporeal video from the operating room was captured via the laparoscopic camera feed, with no audio or room video captured. Multiple surgeons at each hospital contributed cases to this patient series, with trainees participating as both primary and bedside-assistant surgeons. Patients and surgeons analyzed in the study all provided consent, and the study was approved by research and ethics boards from all participating hospitals. All patients undergoing RARP during the data collection period were eligible for inclusion in the analysis.

Data Collection and Analysis

A full description of the variables collected and the video analysis can be found in a previous publication, outlining the construction of the predictive models used here. Patients provided pre-operative urinary function using the 26-item Expanded Prostate Cancer Index Composite (EPIC-26). (Szymanski et al. 2010) Intraoperative variables

were collected by the research team member present in the operating room, including estimated blood loss, degree of nerve-sparing, and the operating surgeon at the various operative steps. Postoperative patient data were collected through chart review, at one year postoperatively to allow maturation of study outcomes. Rating of surgical performance was executed by three trained analysts, with expertise in the assessment methodology, and orientation in the use of the evaluation instruments. The Global Evaluative Assessment of Robotic Skills (GEARS), (Goh et al. 2012) and the Prostatectomy Assessment Competency Evaluation (PACE) (Hussein et al. 2016) were used to quantify the performance of the primary surgeon over six defined steps of the operation: bladder drop (BN, PACE includes the preparation of the prostate), bladder neck dissection (BN), seminal vesicle dissection (SV, PACE includes the posterior dissection), neurovascular bundle and prostatic pedicle dissection (NVB), apical dissection (AD), and the urethrovesical anastomosis (UVA, divided on PACE into needle entry, needle driving, and approximation). Both of these assessment instruments have excellent validity evidence supporting their use in this context. As previously published, raters had strong or excellent interrater reliability metrics, with Cronbach's alpha of 0.73 for GEARS and 0.88 for PACE scores.

Standard Setting Methodology

In this study, we build upon a previously published methodology paper, that used a multivariable model of the skill-outcome relationship in RARP to set performance standards that are patient-specific. (M. G. Goldenberg & Grantcharov 2017) Using Receiver Operating Characteristic (ROC) curves, we determined the optimal trade-off between sensitivity and specificity of the model, which can be adjusted depending on the assessment characteristics and desired pass/fail rate. Reconfiguration of the regression equation allows one to identify the performance score that best predicts the likelihood of a given outcome, based on the chosen probability statistic selected from the ROC curve. This means that the standard is dependent not only on the outcome of interest, but also the patient factors that served as covariates in the model, in the previous paper being age, prostate volume, and BMI. In this study, we applied this

methodology to a prospective cohort of patients, operated on by multiple surgeons of various skill-levels, with two different outcomes, in order to ensure the generalizability of our standard setting method.

Creation of Composite Variables

Our previous methodology paper used three regression models from a retrospective patient cohort to set performance standards. One model used an total GEARS score as the variable of interest, calculated as the mean score of each procedural step. While an appropriate way to provide an overall approximation of the performance of the surgeon over the course of the operation, simply using the mean assumes that all procedural steps are of equal importance to the outcome of interest. In this study, we selected those GEARS or PACE procedural step scores that on bivariate analysis were significantly higher in patients with a favourable outcome, and included them as a variable of interest in a binary regression model with clinically relevant patient covariates. The beta-coefficient in the regression models for each of these procedural steps was used to devise step-weights. In addition, combining these weighted step scores, we created a composite GEARS and/or PACE score that ignored non-significant procedural steps, and appropriately weighted the included steps based on their relation to the outcome of interest. As a final step, composite GEARS and PACE scores were included in binary regression models.

15.3 Results

Description of Patient and Surgeon Cohort

As outlined in our prior manuscript, this cohort consists of 91 patients who underwent RARP during the study period, with 31 surgeons contributing performance data. A variety of case experiences and skill levels were seen across the surgeon cohort, with nearly half of cases being performed by faculty surgeon with more than 250 cases completed prior to the study period. Patient characteristics were typical for a

cohort of RARP procedures, with only one postoperative blood transfusion, and no conversions to open surgery, intensive care admissions, or deaths during the study.

Step Weighting and Composite Variable Models

Although three outcomes were used in the previous study, erectile function is not included in this current analysis, as no individual PACE step scores were significant on bivariate analysis. This analysis uses two regression models with urinary continence at three months postoperatively as the dependant outcome. On bivariate analysis, GEARS scores were higher in continent patients during the BN, NVB, AD, and UVA steps, as well as PACE scores during the BN and two components of the UVA step (needle entry and needle driving). The third model used included PSM as the variable of interest. In patients with positive surgical margins, PACE scores were significantly lower during the SV (seminal vesicle and posterior dissection) and AD steps. These step scores were put into binary regression models with patient covariates selected based on their clinical relationship to the variable of interest, and beta coefficient values are listed in Table-22. The weights assigned to each step in the composite scores is displayed in Figure-5, for the composite GEARS score and PACE score for continence, and the composite PACE score for PSM. Finally, each composite score was included in the binary regression model, and unsurprisingly all three remained independently significant of patient covariates. Forest plots for these three models are seen in Figure-6, and for reference purposes the odds ratio and confidence interval for the overall GEARS and PACE scores used in the previous study are shown in red.

Table-22: Steps in Multivariable Models and Beta Coefficients used in Weighting

	Beta Coefficients	Odds Ratios	95% Confidence Interval		*p*-value
Continence†					
GEARS Scores					
Bladder Neck GEARS	1.07	2.93	1.23	6.97	.015
NVB GEARS	.82	2.26	.90	5.69	.082
Apex GEARS	1.03	2.80	1.05	7.42	.039
UVA GEARS	1.82	6.19	2.07	18.55	.001
PACE Scores					
Bladder Neck PACE	.39	1.47	0.91	2.38	.113
UVA (Needle Entry) PACE	1.22	3.38	1.53	7.49	.003
UVA (Needle Driving) PACE	1.05	2.85	1.40	5.78	.004
Positive Surgical Margins*					
Seminal Vesicles PACE	- .55	.58	.34	.99	.044
Posterior Dissection PACE	- .40	.67	.40	1.14	.138
Apex PACE	- .71	.49	.26	.93	.028

Data extracted from multivariable analysis.
†Continence covariates included patient age, nerve-sparing, prostate volume > 46.5 grams, BMI > 27.5, posterior reconstruction
*Positive surgical margin covariates included tumor stage, Gleason grade, and PSA value at time of surgery

Figure-5: Weights assigned to each step in the composite scores

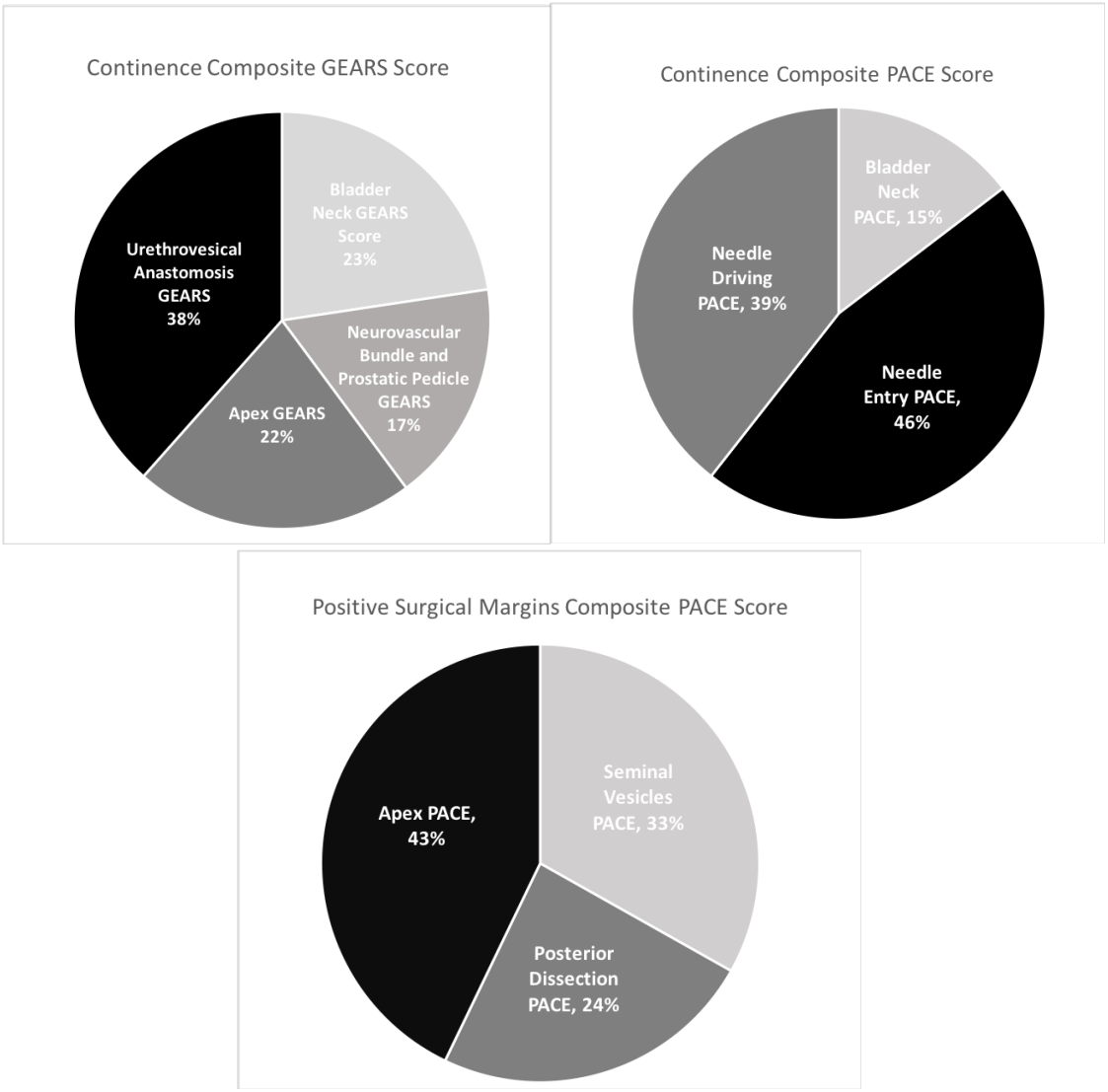
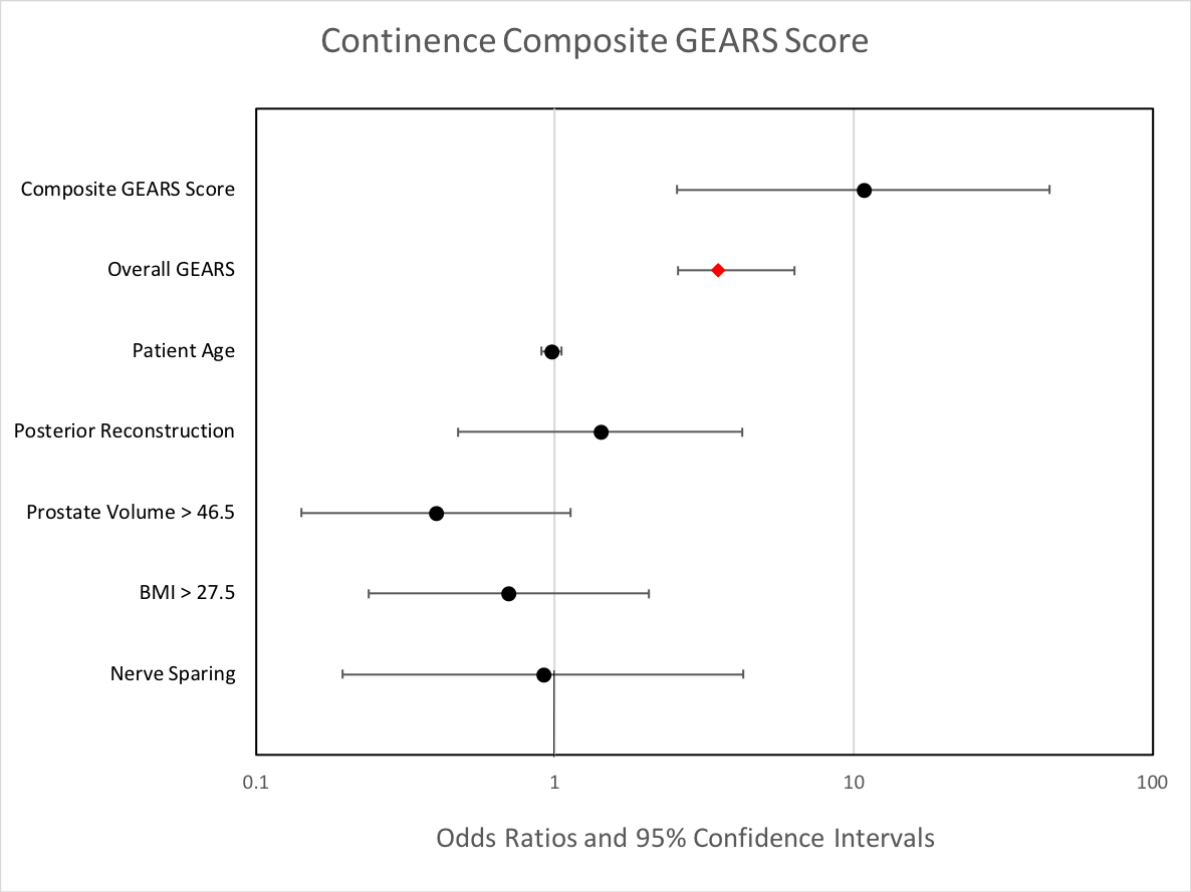
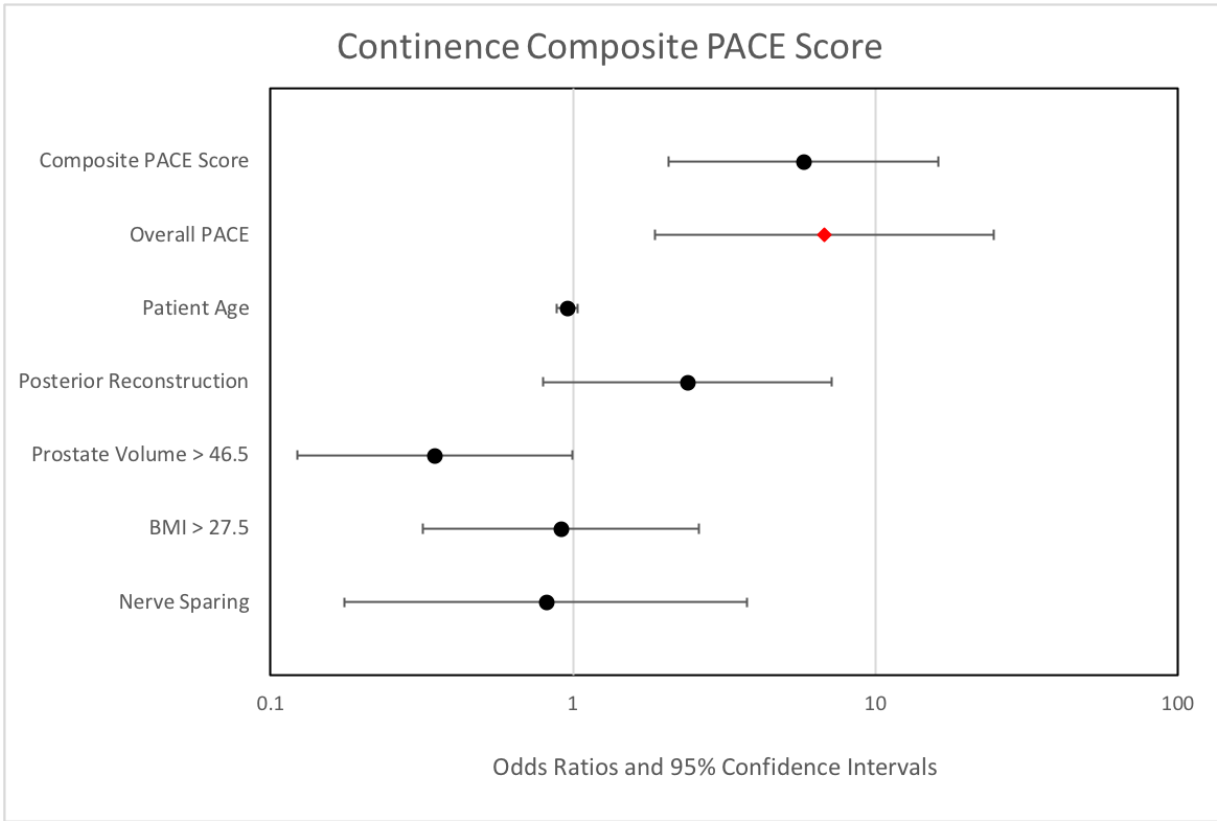


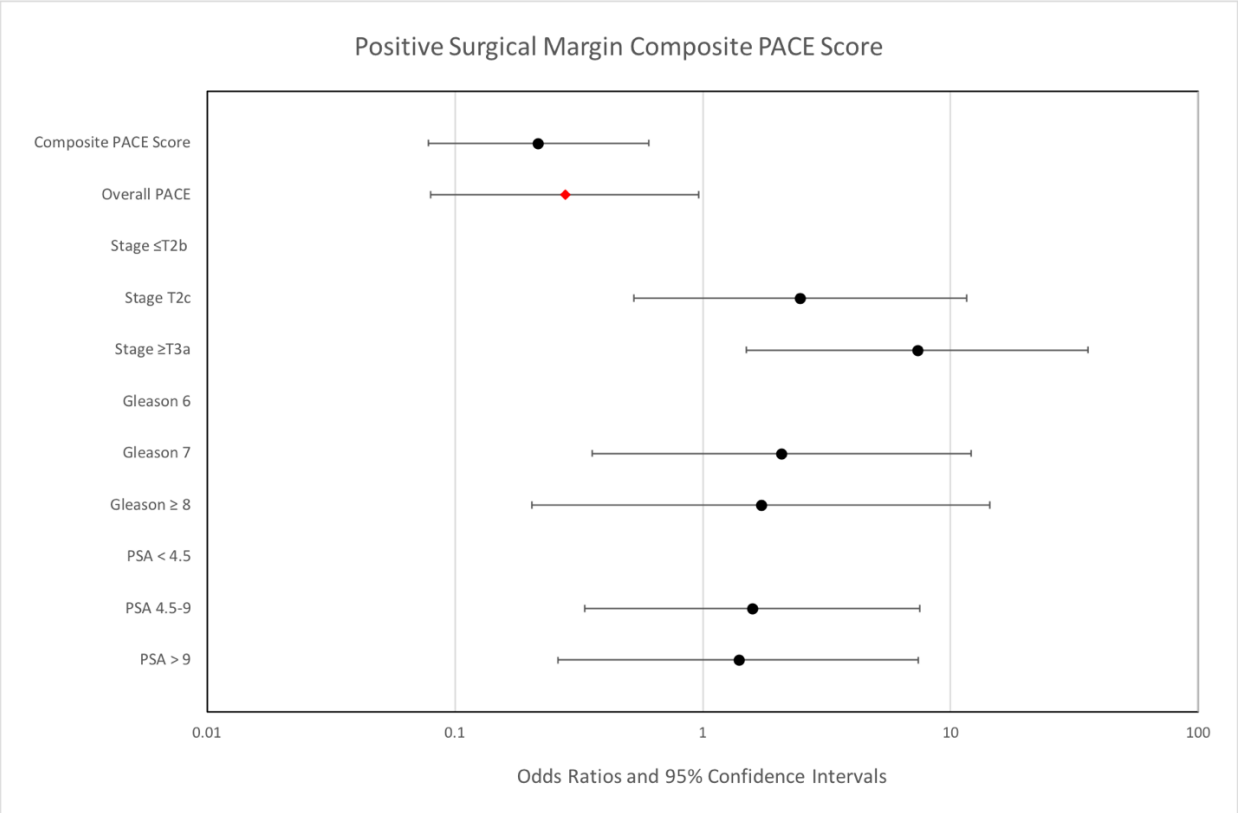
Figure-6: Multivariable Regression Models using Composite Weighted Models



*Overall score (unweighted) shown in red



*Overall score (unweighted) shown in red



*Overall score (unweighted) shown in red

Knowledge Translation and Example

Using the above data, we created a user interface (UI, see appendix-4) that allows our standard setting method to be used by educators or policy makers to determine both an overall standard for the case using the composite measure, or a standard for a given procedural step if more applicable (i.e. assessment of a resident or fellow performing a discrete operative step during training). First, the user inputs the patients parameters included in the multivariable models, in the displayed example a 65 year old patient with a BMI of 27, planned nerve sparing, 60 gram prostate, Gleason 7, T2c disease and a preoperative PSA of 10. Next, the user inputs the probability of each given outcome, in our example a 50% probability of continence at 3 months postoperatively, and a 20% chance of a positive surgical margin. Finally, the approximate scores for each step of the operation are entered, in this example the median scores for each GEARS and PACE score in our cohort for the listed steps.

These can be modified by the user, based either their own median performance scores for steps they plan to complete, or the performance of other trainees participating in the case. The lower half of the interface provides the benchmark GEARS/PACE scores for the two outcomes selected, as well as individual step benchmarks based on the step scores inputted above.

15.4 Discussion

This methodology paper describes the further validation and optimization of a novel method of standard setting for procedural assessment. The predictive properties of objective performance scores can be leveraged to provide benchmarks in surgical performance that account for postoperative outcomes, as well as confounding patient factors. Weighting operative steps based on their association with clinical outcomes not only improves the accuracy of the predictive model, but also allows educators to set standards for individual procedural steps that reflect their importance in determining a successful outcome. Finally, we designed and published an open access platform for surgeon educators to use this methodology in the assessment of RARP technical skill.

Objective rating tools of surgical skill or performance have been used in the operating room for over a decade.(Vassiliou et al. 2005) However, until recently studies using these measures have assessed their effect on training and other educational outcomes, rather than clinically significant ones. As such, our described methodology represents the first time that individual procedural steps have been assigned weighting based on the amount or variation they account for regarding a specific outcome. This statistical approach allows for adjusted benchmarks to be set, both for the overall procedure and the composite operative steps that have influence over a given outcome. This allows for this methodology to applied both to formative and summative intraoperative assessments of trainees completing a single procedural step, as well as for surgeons demonstrating their ability to perform above a minimum level of proficiency, prior to independent practice.

Quality improvement in surgery has mainly focused on implementing strategies to optimize patients preoperatively, (Cui et al. 2017) and improve or shorten the postoperative course of recovery. (Azhar et al. 2016) However, recent investigation into the influence of intraoperative performance on patient safety and outcomes has brought issues of surgical training and accreditation to the forefront of QI research. (Dimick & Varban 2015) Objective measures of surgical skill, including video-review of operative footage, have been proposed as metrics in privileging pathways for surgeons that go beyond simple measures such as previous training and case experience.(Tam et al. 2017) However, as with any metric of quality, benchmarking is crucial in the credentialing process, and ideally these benchmarks have some clinical implication that are supported by evidence they improve the safety or well-being of patients. (Panzer et al. 2013).

Although this methodology was created with a rigorous statistical approach, there are limitations that must be acknowledged when interpreting these procedural standards. First, although we used established tools for evaluating technical skill, these instruments do not capture the non-technical skills of the surgical team. Teamwork and communication in the operating room likely contribute to the success of an operation, despite the literature not yet supporting this in many surgical approaches. Secondly, it must be stated that although the predictive models used in this study can predict the chosen outcomes reasonably well, there remains a degree of uncertainty associated with the trade-off between sensitivity and specificity that must be acknowledged when making high-stakes decisions with this standard setting method. These limitations both speak to a wider issue, that the variability in surgical outcomes cannot be accounted for with the variables we currently use to predict an operation's success. In addition to non-technical skills, anatomical complexity, as well as human and system factors remain difficult to quantify, and may themselves contribute to postoperative outcomes.

15.5 Conclusion

This study represents the refinement of a novel approach to standard setting that uses predictive models to create benchmarks in technical performance based on patient outcomes. Weighting performance scores on individual operative steps based on their association with outcomes of interest allows educators to understand minimum standards for use in both formative assessments in training, and summative assessments in accreditation or privileging practices. Creating a UI based on this work allows for more effective knowledge translation and broad implementation of this methodology.

16 Discussion

16.1 Thesis Synthesis

This thesis has examined the relationships between assessments of technical skill in the robotic surgery operating room, and measures of patient outcome, and leveraged this relationship to set benchmarks in performance that focus on patient safety.

Beginning with an exploration of the current methods of both assessing robotic technical skill in urology identified those evaluation methods and tools that are supported by robust validity evidence, for use in both formative and summative decision making. Although novel methods for doing so are emerging, it seems that GRS-based assessments remain the most accurate and reliable method of stratifying trainee and staff surgeons by technical ability.

Traditional, educational standard setting methods have been used to successfully set pass marks in assessments of health care professionals for many years. There have been examples in recent literature that both item-centred and participant-centred techniques can be used to evaluate a trainee's technical skills, in

both the simulation and clinical environments. Operating room-based, GRS assessments can accurately classify surgical residents by their level of training and experience, when performing even laparoscopic surgical procedures. This data supports these assessment's use for providing skill level-appropriate feedback and directing trainee learning. However, while this *relationships to other variables* evidence is crucial in supporting pass/fail decisions, we do not yet know how these decisions may impact patient outcomes and surgical safety. To make truly informed high-stakes decisions regarding whether a surgeon or trainee is safe to carry out a surgical task in the clinical environment, the *consequences* of surpassing or failing to reach a pre-determined score threshold is essential.

The increased use of RAS in urology and the demand for robust, outcome-based assessments during the implementation of CBME had identified a significant knowledge gap addressed in this work. Using a carefully designed, retrospective matched-cohort study, the relationship between GEARS scores and patient outcomes was investigated. The technical performance of a single surgeon was evaluated, and compared between patients with or without urinary continence in the early postoperative period. Using this patient-centred outcome, and an objective GRS as the variable of interest, an independently predictive relationship was found. At the time of publication, this conclusion was truly novel in the field of RARP, and subsequently these findings have been replicated in studies published by other investigators. This finding added two important pieces of validity evidence to this assessment method. First, it indicates that this score may be able to not only differentiate between different trainees or surgeons of different experience of training level, it may also be able to capture important variations in performance of a *single* surgeon. While this finding must be interpreted in the context of a single surgeon, retrospective study, this highlights an important concept in medical education. While studies of technical skill acquisition have shown that this occurs over a set trajectory or pattern, fluctuations in performance from one procedure to the next may be significant enough to impact patient outcomes. Certainly, non-technical skills such as inter-team communication, leadership, and decision-making have been shown to have a direct association with technical skill scores. Additionally, patient (anatomy, disease

characteristics) and system-level factors, while often difficult to adequately capture, may have a direct effect on a surgeon's ability to successfully carry out a task or procedural step in the operating room, and therefore warrant future investigation.

Based on the findings of our original study, we hypothesized that this predictive relationship between performance and clinical outcome could be harnessed to set performance standards that distinguish between competent and incompetent surgical care. The multivariable models produced in the retrospective analysis provided an idea of the impact that surgical skill has on the variability in outcomes following RARP, particularly when placed in the context of clinically relevant patient factors. This regression analysis provided coefficients that were used to assign various weightings to these included independent variables. Providing values representative of a patient's clinical characteristics (i.e. age and BMI), a minimum performance score is determined that must be exceeded in order to provide the patient with a given probability of experiencing the outcome of interest (in this study, urinary incontinence). This methodology is entirely novel, and uses a statistical approach to bridge an evidence gap in surgical education.

Although these two studies provide innovative clinical data and educational techniques to the literature, they use a limited patient sample and study design. To prospectively validate the findings of the initial retrospective study, a multicentre, multi-surgeon and multi-institution study was undertaken. A large cohort of patients were enrolled across three hospital sites over a nine-month period, and over 30 surgeon participants provided performance data from procedural steps of RARP. Through a stepwise approach, our initial findings were replicated using multivariable and sensitivity analysis, and internally validated using a traditional predictive modelling method. This study further elucidates the relationship between not only surgeon technical skill in RARP and urinary function, but also sexual function and even PSM occurrence. All three of the selected outcome have important implications for patient well-being and cancer control, providing additional and important validity evidence for the scoring tools used, GEARS and PACE.

Using prospectively collected data from across multiple surgeons and hospitals allowed for further refinement of this standard setting method. Using the statistical method described in earlier work, a method of determining benchmark performance scores across three distinct clinical endpoints was created. Furthermore, the use of surgical step-weighting in the predictive models improved the ability of the model to accurately predict these clinical outcomes. This work allows for surgeons to provide patients with predictions around their postoperative outcomes based not only on their individual risk factors, but also incorporating their own technical performance. Additionally, the methodology allows for educators and credentialing bodies to set benchmarks in technical ability that are not arbitrary or based on consensus, but rather use real-world data to set standards that have true clinical meaning.

16.2 Assessments of Technical Skill in the Clinical Environment

The role of assessment in the clinical environment, whether focused on technical skill or otherwise, is undergoing a cataclysmic shift in an outcome-focused, competency-based education system. The use of formative and iterative assessment strategies has created an unmet demand for readily digestible, yet methodologically robust measures of trainee competency, with an early focus centering around technical clinical skills needed for effective patient care. Learners demand more oversight and guidance in directing and personalizing their training, while accreditors continue to seek out novel ways of ensuring that the needs and demands of patients are met through standardized and safe patient care.

This work in particular feeds into the concept of entrustability. Setting standards that distinguish good from poor patient outcomes allows assessors to better identify when a trainee has reached a minimum level of entrustment, and provides validity to high-stakes decision making about a resident surgeon's ability to operate independently. Benchmarking the level of skill at which a trainee can operate independently, whether an operative step or an entire procedure, allows stakeholders

with quantifiable and concreate degrees of skill at which a trainee can be trusted to operate without supervision, and can provide junior faculty or even trainee-peers with objective information with which to help make these judgments. EPA frameworks should therefore benefit greatly from the described methodology in this thesis, and this will create more acceptance of competency-based curricula in surgical education.

Standard setting of trainee skill using traditional methods relies heavily on expert consensus around the performance of a borderline trainee, and therefore the conceptualization of such a trainee in a specific context. While a valid process in setting performance standards in procedural skill assessment, significant biases may affect the outcome of these processes when applied to clinical training in the real-world. This thesis work highlights one such barrier, the influence of patient and disease factors when assessing surgical residents in the operating room. Unlike SBA's, the clinical environment is laden with inconsistencies when comparing evaluations of skill from one case to the next. When directly observing residents in the operating room, factors such as surgical complexity, surgeon stress or anxiety, and extraneous time and resource pressures may play a role in determining the score or feedback provided based on a trainee's performance. Many of these factors are difficult or impossible to control for in the context of skills assessment, and this creates a challenge for program directors and other stakeholders when comparing a cohort of trainees to one another. Benchmarks should be a reflection of the purpose of the assessment, while adjusting for as many of these factors as possible. This allows a level playing field when comparing trainees at graduated levels of skill across the curriculum and increases the validity of such high-stakes decisions.

16.3 Technical Performance and Patient Outcomes

The mounting evidence for the role of surgeon or physician technical performance in determining patient outcomes is forcing healthcare organizations to examine more closely how their employees' function as individuals, and in team-based activity. Although the outcome of a given patient's experience in the healthcare system

cannot be totally pinned on their surgeon, clear differences in even basic technical skills do exist. Furthermore, the variation within a given surgeon's performance from one operation to the next must be acknowledged as a significant influencer of surgical outcomes, and the work in this thesis and across the surgical literature has shown that system and team-level factors likely play a key role in determining how a surgeon will perform on a given day, in a given case, or even in a given operative step.

Although the seminal work by Birkmeyer et al is most often cited as direct evidence for this phenomenon, the literature is filled with multiple examples of the interplay between technical skill and outcomes. Fecso et al compiled a systematic review of the evidence examining this exact topic, and demonstrated that studies have looked at this relationship from multiple angles, using different metrics of performance and outcome to add to the overall pool of evidence. Even in as specific an area as radical prostatectomy, multiple studies have consistently shown that technical performance explains part of the variation in important, patient-centered outcomes, such as urinary continence. This postoperative symptom was chosen as our initial primary outcome because it has shown to significantly impact patient quality of life. Unlike some bariatric and oncological procedures in other fields, radical prostatectomy is overall a safe and fairly well tolerated operation. However, the 'side effects' of this treatment can carry significant implications for the psychological well-being of patients undergoing this procedure. Stress incontinence, when bothersome to a patient, can have social, psychological, and even sexual consequences for men, and as physicians our first goal should be to try and minimize the number of men experiencing this symptom in the postoperative period. However, most urologists and patients would likely agree that oncological success remains far and away the most important postoperative outcome. The evidence presented here demonstrates that PSM, although clearly in part a product of biological, disease factors, remains independently associated with the surgeon's technical execution of the surgical steps. Again, this finding is not in isolation or opposition of the surgical literature, with evidence from over 10 years ago in laparoscopic radical prostatectomy demonstrating a link between technical execution of the procedure and PSM in a limited cohort of patients.

These findings have ramifications for not only the education of future surgeons in the competency-based curriculum, but also for practicing surgeons performing radical prostatectomy in both the academic and community setting. Currently there are no limitations restricting or regulating surgeons' ability to carry out a procedure, despite many being 20 or more years removed from training or assessment in that given procedure. This standard of surgical credentialing is woefully outdated when compared to other high-reliability industries, most of which do not even rely as heavily on human performance in order to achieve good outcomes. In aviation, pilots are regularly assessed during their careers, and must demonstrate not only adequate knowledge, but also technical proficiency when executing routine and emergency maneuvers. This strict regulation of both technical and non-technical skill exists in an industry where increasingly it is the aircraft and on-board computer systems that control and carry out these important processes. In opposition to this, humans remain at the very heart of most if not all functions within the operating room environment, and poor performance or execution of individual or team-based skills can have enormous consequences for the patient on the operating room table. Surrogates of surgical quality such as surgeon volume may not be a direct enough means of credentialing surgeons, as this would prevent low-volume but technically competent surgeons from carrying out operation. In the context of Canadian healthcare this could mean that patient care is centralized to high volume hospitals at the expense of close to home, high quality surgical care. Only by including high-stakes assessments of technical skills into credentialing practices can we ensure that technical competency, at minimum, dictates whether a surgeon is safe to independently operate on his or her patients.

16.4 Improving Assessment Validity by Benchmarking Performance with Patient Outcomes

As demonstrated in this thesis, current research has focused on increasing the variety of assessment tools in surgical education, with a lack of focus on the evidence support the use of these instruments in determining the competency of a trainee.

Authors have continued to use the outdated approach to assessment validation, claiming that their approach to technical or non-technical skills assessment is 'valid' based on single studies of specific groups of trainees or surgeons. Claiming that a given evaluation is valid across multiple contexts, especially for determining something as complex as trainee competency, is dangerous when implementing these assessments in training curricula. As modern approaches to validation have stated, the often high-stakes nature of medical education demands that assessment tools can contribute to defensible, reproducible, accurate, and reliable decisions regarding a trainee's ability to function as a safe and independent practitioner. It is irresponsible for program directors or accrediting bodies to claim that their program of assessment ensures competent surgeons, without exploring the consequences of these assessments on patient outcomes and safety.

Exploring the consequences of both the types of assessment and the assessment tools themselves has been at the core of this thesis work. Systematic examination of the existing literature assessing the technical skills of the robotic surgeon demonstrated that although the literature uses an evidence-based approach to creating and evaluating the *educational* significance of their instrument's scores, in most cases they do not adequately explore the *clinical* significance of these scores. Benchmarking performance using these assessments is the first step in the process of incorporating these instruments into summative decision making, and little to no studies have done this using conventional or novel methods. This lack of exploration of standard setting in RAS assessments is further highlighted by our systematic review of these methodologies as applied across procedural, technical skill assessments. Participant and item-centred methods of setting score benchmarks exist and have been successfully applied across the surgical education literature, from the simulation lab to the operating room. Despite the described limitations of these existing methods, application of these strategies in the urologic and RAS literature would at least represent a step toward the incorporation of technical skill assessments into high-stakes decision making, for use in a CBME curricula.

Not only does this work represent an early example of benchmarking performance in RAS, it also provides valuable consequences evidence for these standards. Taking a novel approach to creating these standards that is rooted in patient outcomes allows increased confidence amongst educators and credentialing societies when making decisions regarding privileging of surgeons performing RAS. The use of expert consensus alone when determining whether a trainee has reached a level of 'competence' to independently perform a potentially life-altering operating will not be long tolerated in an increasingly evidence-driven society. As patients become more informed regarding the impact that the physician themselves has on their outcomes, they will demand that their surgeon is being continuously held to a standard that provides them some semblance of security and safety.

17 Limitations

17.1 Challenges of WBA in Surgery

This work has added important validity evidence for the use of GRS assessments in the clinical world, through exploration of the association between performance scores and clinical outcomes. The importance of assessing technical performance in the operating room cannot be stressed enough, as it remains nearly impossible to quantify the direct impact that these aspects of surgical training have on patient safety in other assessment environments. However, certain limitations that exist in the real world that may temper the use of WBAs in a competency-based curricula. These include time and resource restrictions, issues around generalizability of assessment scores, and attribution bias regarding the performance-outcome relationship.

17.1.1 Time and Resources

Surgical training must balance service and education, and maintaining this delicate equilibrium can serve as a barrier to the implementation of assessment

strategies in residency programs. While essential, programs evaluating trainee performance demand time, resources, and energy to implement successfully. Educators and mentors must be willing to provide both ratings of skill and elements of directed feedback in order to CBME to be effective, and unlike other aspects of academic medicine, physicians may not be compensated for these extra tasks. Fortunately, most surgeons at academic institutions understand the importance of residency education, and this buy-in has allowed for surgical training to continue to thrive internationally. However, the increasing rigour and frequency in which skills assessments must take place in an outcome-based curriculum may strain the time and resources of residency programs, making it imperative that these issues are addressed early in the rollout of competency-based surgical education.

Time is a precious resource and may limit the adoption of iterative evaluations of technical surgical skill. Increasing numbers of healthcare users increase the amount of clinical demand placed on physicians, and this further directs our time and energy away from education and toward patient care. This increase in workload has been accompanied at academic institutions by a demand from accreditation institutions for more focus on faculty development and investment into developing more sophisticated educational systems and curricula (Nousiainen et al. 2017). Beyond the time required to provide effective evaluation and feedback, CBME requires faculty participation in remediation of borderline trainees, review of existing curricula, and other educational committee membership. Without changes at the culture-level in our current healthcare environment, these growing demands for educational activities will be met with resistance, especially given the lack of monetary incentivization (Caverzagie et al. 2017).

17.1.2 Generalizability

An obvious limitation of this work that must be addressed is the lack of generalizability outside of procedural, technical skills assessment. In the ever broadening world of competency-based assessment, the use of a predictive model to

set performance standards truly only applies to a portion of assessments that are required of a trainee in order to deem them competent in their field of specialization. The broad definitions of competency in clinical medicine and surgery involves the evaluation of all domains of practice, many of which cannot be as easily scored using traditional, psychometric-based approaches (FRANK 2005). As such, a huge amount of academic and translational work is still required before this model of patient outcome-based standards can be applied across the gamut of competency-based assessments. As of now, the vast majority of educational literature assessing the consequences evidence of assessment tools comes from the Medical Expert domain, in particular procedural skill evaluation. However, with the broadening of frameworks to make collecting validity evidence simpler for qualitative assessments, I believe we will see more evidence linking domains such as professionalism and scholarship to patient-centered clinical outcomes. (Harris et al. 2017).

Importantly, one must address this issue from a psychometric perspective. First, one should determine the number of assessments needed to make a reliable and accurate judgement of one's competency in procedural skill. Reed Williams and his colleagues at Southern Illinois University have studied this question extensively, and their data indicates that a person needs to be assessed 17-23 times in order to produce reliable ratings of technical skill (Williams et al. 2017). Others have looked at this same question across multiple contexts. Both Williams (Williams et al. 2015) and Gofton et al (W. T. Gofton et al. 2012) found that in order to achieve an interrater reliability of 80%, one must assess a trainee performing a technical task a minimum of 5 times. Beard and colleagues in the UK also found through a generalizability study that 5 assessments were required to provide reliable ratings of trainee technical skill. Conversely, Crossley reported that 6-8 assessments are required to provide generalizable ratings of non-technical skill in the operating room, using the NOTSS instrument.

In addition to accumulating a minimum number of assessments per trainee in order to provide accurate ratings of skill, it is important to use an adequate number of assessors to limit the bias associated with observational assessment scores. Dr.

Williams' group recommends that a minimum of 10 raters be used when assessing procedural skill. In a study of clinically-based assessments, Gingrich and colleagues found that raters observations and judgments can be grouped into 5 distinct perspectives, and this finding lends itself to the notion that two raters may observe a performance from a drastically different point of view. Lockyear et al argue that this categorization of rater perspectives reflects the inability for quantitative assessments to adequately capture the factors that lead to a given judgement of an assessor. They conclude that incorporating qualitative components into an assessment strategy may help limit this by allowing raters to more readily justify their decision-making around competency or entrustability.

Finally, the frequency in which trainees are observed may have an impact on the validity of scores generated by these assessments. William's study from 2012 indicated that residents should also be evaluated at least two times per month in the operating room in order to make reliable high-stakes decisions regarding their operative competency (Williams et al. 2012).

17.1.3 Qualitative Assessments in Surgical Education

In an outcome-based assessment program, decisions around competency should use multiple styles of assessment technique, to provide stakeholders with a wealth of diverse performance data for a given trainee. While the techniques used in this thesis focus primarily on quantitative methods of assessment, it is important to underscore the importance of qualitative feedback and assessment methods in a surgical training program. Quantitative assessment tools are commonly used because educators are traditionally more comfortable with making pass/fail decisions based on numerical data. They align well with the scientific principles that most physicians are familiar, using psychometrics and inferential statistical methods to draw conclusions about trainee performance. However, while essential to the identification of those trainees who require remediation, these methods alone may not adequately explore the

underlying constructs that have prevented some from progressing while allowing others to excel.

The methods used in this thesis are purely quantitative. They harness Likert-type scales to create a numerical representation of surgical performance that can be readily inputted into statistical models predicting the outcome of a patient undergoing RARP. However, these findings are limited by their ability to give tangible feedback to the clinicians involved. Surgeons crave specific, targeted feedback on the various aspects of their technical skill and approach to a procedural step. A score is only as valuable as what it represents, and numerical differences between a well and poorly executed surgical technique may not be readily translatable into obvious feedback for learning and improved performance.

Modern approaches to surgical quality improvement have found a way to harness qualitative methodologies into their frameworks. Coaching is a prime example of the translation of performance scores into digestible and addressable targets for improved patient care. Work from Greenberg (Greenberg et al. 2015), Hu (Hu et al. 2016), and others have demonstrated the role of structured coaching frameworks in the broader field of surgical quality improvement. Using teaching strategies that have been well-developed in other industries, these authors have provided us with a method of actioning the valuable quantitative assessment data that we have been and will continue to collect, into individualized and targeted feedback for surgeons.

17.1.4 Attribution Bias and Unexplored Consequences

The standard setting method described in this work is based on the fundamental relationship between a surgeon's technical execution of a procedural task or step, and the postoperative outcome of the patient. This assumption relies on this relationship being one of causation, rather than association, and must be addressed as a key limitation. Although the models used to create the standard include patient factors and attempt to address random and fixed effects through robust validation techniques, it is

impossible to account for all unmeasured confounders in this analysis. This leaves the door open to misinterpretation of these data, in that patients or other stakeholders may not understand this nuance and rely too much on the quantitative standards that are calculated with this method. Non-operative, non-surgeon, and non-patient factors certainly play a role in dictating many postoperative outcomes, and further work is needed to understand how best to quantify these.

There are implications for this work that must be addressed as potential limitations to its dissemination as a tool for accreditation. As an example, surgeons may not be willing to accept that their technical skills should be put under scrutiny as long as their operative outcomes are satisfactory. Implementing these standards into credentialing practices implies that failure to meet benchmarks in performance would lead to a limitation or withdrawal of operating privileges for a given procedure. While in some or most cases this will have a net benefit on patient care, there are ramifications to limiting the number of surgeons able to carry out certain procedures. Increasing the operative volume of a fewer number of surgeons could increase wait times in an already stretched Canadian system. Furthermore, patients may be forced to travel to receive certain surgical care, no longer available in their community due to these restrictions of practice. However, this centralization of care may already be occurring due to limitations on resources and growing sub-specialization through fellowship training.

17.2 Controversies Around Robotic Surgery in Canada

Socialized medicine in Canada, like other healthcare systems around the world, has strengths and weaknesses. While free, public access to essential healthcare provides Canadians with a sense of national pride, it comes at a price, both from a government expenditure perspective, and from inherent limitations in the types of treatments and technology available. Robotic surgery has been under scrutiny since its arrival on the healthcare scene, and its expense and utility have been brought into question by a number of stakeholder groups, including Health Quality Ontario (Health Quality Ontario 2017). This Health Technology Assessment (HTA) was completed in

2017 by a government-appointed taskforce of physicians and recommended against public funding of robotic surgical programs. They cited primarily a lack of high-level evidence supporting its benefit for patient outcomes, and a cost-utility analysis that showed significantly higher associated costs. This recommendation has sparked a lively debate among the public and healthcare policy-makers alike, and ongoing lobbying of the provincial government in Ontario has so far halted any final decisions from those bodies responsible for healthcare spending. However, it must be acknowledged that this debate on the best implementation of robotic surgical technology in our healthcare system continues, as this has the potential to limit the use of our research in our own province and country.

17.3 Non-Technical Skill Assessments

Not included in these studies are assessments of surgeon non-technical skill, and the omission of these factors must be included as a limitation of the work. Non-technical skills, such as communication, teamwork, and leadership, play an important role in the overall performance of surgical teams. Fecso et al recently published their series that looked at the temporal relationship between safety events, technical performance, and non-technical skills in bariatric surgery, finding a strong association between these factors in the operating room (Fecso, Kuzulugil, et al. 2018). A systematic review showed that non-technical skills have multiple correlates in the analysis of surgical safety events, and despite the mixed quality of these data, they concluded that focusing on teamwork interventions and other non-technical skill development strategies will have a positive impact on overall performance and patient safety (Gjeraa et al. 2016). Additionally, evidence supports the direct correlation between technical and non-technical performance (Mishra et al. 2007; Brunckhorst et al. 2015). Importantly, work has shown that these non-technical skills are trainable. Steven Yule, lead investigator in the creation of the Non-Technical Skills for Surgeons (NOTSS), provided level-one evidence that surgical trainees can adopt the core principles of good communication, leadership, teamwork and situational awareness over a short training period (Yule et al.

2015). This finding has been replicated across multiple contexts (Pena et al. 2015; Dedy et al. 2015).

18 Future Directions

18.1 Expanding the Methodology

This thesis examined the association between surgeon performance and patient outcomes and devised a means of using this relationship to set evidence-based standards in technical skill for use in educational and accreditation activities. As mentioned above, the procedure and instrument-specific nature of this methodology limits its direct generalizability across other areas of surgical education. Additionally, multiple facets of performance in the operating room are not included in this model, and this forms the basis for future work in this field.

Multiple methods of capturing non-technical skills have been explored in the literature, that allow for both the measurement and quantification of factors such as situational awareness, leadership, and decision making (Yule et al. 2015). Objective assessment of these factors is difficult, as live-assessment may be biased without blinding, and analyzing interactions involving multiple team members is difficult without the ability to re-watch certain moments within a case. However, novel methods of video-capture in the operating room may provide a means of prospectively collecting and evaluating non-technical skills in surgery, allowing these factors to be more readily studied in association with patient outcomes (M. G. Goldenberg, Jung, et al. 2017).

An area of investigation recently adopted from the world of engineering and other high-reliability organizations is ‘human factors’ research (Reason 1995). Recent efforts in patient safety have moved away from what is known as ‘Safety 1’, or the analysis of system and human actions that lead to patient harm, to ‘Safety 2’, the identification of human resiliencies that prevent or mitigate harm from occurring in the first instance (Jeffcott et al. 2009). This shift in focus away from a reactive approach to patient safety improvement to a more proactive mentality, has created a new way of assessing these

factors in the operating room (Hu et al. 2012; Ferroli et al. 2012). Further investigation into the generalizability and reproducibility of these frameworks is needed before they can be incorporated into surgical curricula.

Automation of technical skills assessment may provide unique insights into how performance in surgery relates to outcomes. Autonomous rating of surgical video will allow for high volume rating of operative data and integration of these metrics into population-level databases. The use of kinematics and computer vision analysis to track surgical instrument motion, hand motion, and muscle contraction, have led to breakthroughs in our understanding of the effect of these basic motor functions on surgical performance (Azari et al. 2017; Snaineh & Seales 2015). These innovative methods of data capture rely on sensor tracking or computer learning technology to quantify technical skill metrics in an automated manner. This not only cuts down on the time needed to analyze performance, but also limits or nullifies the subjective bias of human raters and allows for real-time scoring of technical skill. This opens the door for future investigation into the use of 'predictive analytics' in performance assessment, that is, the ability to predict whether a combination or pattern of technical events has the potential to lead to human error or patient harm.

18.2 Exploring the Methodology in a Program of Assessment

This described standard setting technique is optimally suited for application to competency-based systems of training and credentialing. The primary goal of this approach to benchmarking surgical performance is to provide important construct validity to the assessments that are being adapted for use in high-stakes assessments. No outcome is more important than patient safety, and therefore it is of paramount importance that educators and stakeholders in surgeon credentialing are confident in the abilities of those they chose to accredit. Without this lens on performance assessment in surgery, we cannot be sure that graduating surgeons entering independent practice are truly competent 'proceduralists,' and must make these

important decisions based on often unstructured evaluation by that individual's mentors. The current process is likely fraught with issues of assessor bias, and the current wide gaps in patient postoperative outcomes across institutions and surgeons with similar practice patterns reflects this outdated system of physicians credentialing.

We envision this methodology being used in CME curricula, across all medical specialties. Evidence supporting the validity of ability or performance assessments across all specialties will embolden stakeholders in recertification and professional development to ensure that physicians are maintaining these skills, through assessments of knowledge, clinical judgement, and procedural performance. Rather than our current credit-based system of CME, we can use this methodology to create targeted programs of assessment and education for practicing physicians to improve their skills and in doing so, improve the safety and outcomes of their patients.

Finally, the process of proctoring surgeons adapting to a new surgical procedure or technology remains unregulated and under the control of device manufacturers. The current model of proctorship involves the observation of a surgeon for a given period of time, arbitrarily determined by the company whose device is being used. This methodology represents an opportunity to improve how we proctor surgeons in this context, through direct observation and objective assessment. Benchmarking technical skill against patient outcomes will allow device companies and hospital stakeholders to have quantitative data supporting the readiness of a surgeon for independent execution of a procedure, in a manner that does not compromise patient outcomes.

18.3 Other Implications of the Methodology

This work represents a small part of a larger shift in how we train, evaluate, and credential within surgery. The merging of quality improvement, education, and patient safety has been spurred in part by their common goal of improving outcomes across the healthcare system. While this research focuses on a specific area of surgical practice,

the methodology used here may have wider implications as part of an ongoing shift toward quality-based accreditation and remuneration.

Using technical or non-technical performance as measures of quality may have broader consequences than were explored in this research. Moving to a quality-based payment and resource allocation system in Canada underscores the importance of using valid and reliable methods of quantifying quality. While procedural volume and patient outcome may have been proposed for use in this context, caution must be exercised when interpreting these variables, without careful consideration of confounding factors, for example, patient clinicopathological characteristics. In this work, we propose the use of physician performance as a measure of quality, as derived from the direct observation of him/her carrying out a clinical task. Similarly, context is of upmost importance when interpreting these data for this purpose. As with all quality indicators, stakeholders must understand that these variables form only a small portion of the variability in determining successful patient outcomes. However, by using a metric that comes from direct and objective observation and evaluation of a surgeon's performance in clinical care, it may be less subjected to the biases that are typically associated with quality measures.

Other sequalae may arise from the use of performance data. Demonstrating that broad differences may exist in the technical or non-technical performance of surgeons, and that these differences are clinically significant, may force stakeholders at the hospital and government-level to move toward centralization of certain procedures or processes. If performance in the operating room is to be used as a means of benchmarking or ranking physicians within a healthcare system, then it may force the hand of those ultimately responsible for patient well-being to redistribute care pathways in a manner that best serves the population. The introduction of expensive healthcare innovation and technology has similarly begun to inadvertently create so-called '*centres of excellence'* in certain procedures or disease states, with patients often travelling within their province in order to have access to a certain procedure. While this has indirectly created a move toward centralization, allocating resource and funding to those

centres with superior outcomes and physician care may more directly shift high-risk or skill-demanding procedures away from smaller institutions into more academic or high-volume ones. While many argue this is an appropriate response to a discrepancy in patient outcomes and wide variability in quality of care, it may also be perceived as detrimental to patients who live in remote regions without ready access to these tertiary or quaternary care institutions.

Finally, we foresee this methodology being useful in the benchmarking of surgical costs for a given procedure, especially in a single-payer healthcare climate as ours in Canada. Similar to setting a standard of surgical quality in relation to patient outcome, the same can be done to determine what level of performance translates to satisfactory levels of cost for a given intervention. Forthcoming work from our group has demonstrated a statistically significant relationship between intraoperative performance and both direct and indirect costs, and this correlation can be similarly leveraged to better standardize and even centralize care to maximize cost-effectiveness.

19 References

Abraham, N.E. et al., 2010. Patient centered outcomes in prostate cancer treatment: predictors of satisfaction up to 2 years after open radical retropubic prostatectomy. *The Journal of Urology*, 184(5), pp.1977–1981.

Admiraal, W. et al., 2011. Assessment of teacher competence using video portfolios: Reliability, construct validity, and consequential validity. *Teaching and Teacher Education*, 27(6), pp.1019–1028.

Aggarwal, R. & Darzi, A., 2011. Innovation in surgical education – A driver for change. *The Surgeon*, 9, pp. S30–S31.

Aggarwal, R. et al., 2008. Toward Feasible, Valid, and Reliable Video-Based Assessments of Technical Surgical Skills in the Operating Room. *Annals of Surgery*, 247(2), pp.372–379.

Aghazadeh, M.A. et al., 2015. External validation of Global Evaluative Assessment of Robotic Skills (GEARS). *Surgical Endoscopy And Other Interventional Techniques*, 29(11), pp.3261–3266.

Aghazadeh, M.A. et al., 2016. Performance of robotic simulated skills tasks is positively associated with clinical robotic surgical performance. *BJU International*, 118(3), pp.475–481.

Ahmed, Maria et al., 2013. Actual vs perceived performance debriefing in surgery: practice far from perfect. *American journal of surgery*, 205(4), pp.434–440.

Ahmed, Najma et al., 2014. A Systematic Review of the Effects of Resident Duty Hour Restrictions in Surgery: Impact on Resident Wellness, Training, and Patient Outcomes. *Annals of Surgery*, 259(6), pp.1041–1053.

Albanese, M.A. et al., 2008. Defining characteristics of educational competencies. *Medical education*, 42(3), pp.248–255.

Alemozaffar, M. et al., 2015. Benchmarks for operative outcomes of robotic and open radical prostatectomy: results from the Health Professionals Follow-up Study. *European Urology*, 67(3), pp.432–438.

Alemozaffar, M. et al., 2014. Validation of a Novel, Tissue-Based Simulator for Robot-Assisted Radical Prostatectomy. *Journal of Endourology*, 28(8), pp.995–1000.

Alzahrani, T. et al., 2013. Validation of the da Vinci Surgical Skill Simulator across three surgical disciplines: A pilot study. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 7(7-8), pp.E520–9.

America, C.O.Q.O.H.C.I.Institute of Medicine, 2001. *Crossing the Quality Chasm:*, National Academies Press.

Amirian, M.J. et al., 2014. Surgical suturing training with virtual reality simulation versus dry lab practice: an evaluation of performance improvement, content, and face validity. *Journal of Robotic Surgery*, 8(4), pp.329–335.

Hung A. et al, 2011. Face, Content and Construct Validity of a Novel Robotic Surgery Simulator. *JURO*, 186(3), pp.1019–1025.

Leverage M., 2015. While you slept: bad behaviour and recording in the operating room - BJUI. Available at: http://www.bjuinternational.com/bjui-blog/while-you-slept-bad-behaviour-and-recording-in-the-operating-room/.

Arain, N.A. et al., 2012. Comprehensive proficiency-based inanimate training for robotic surgery: reliability, feasibility, and educational benefit. *Surgical Endoscopy And Other Interventional Techniques*, 26(10), pp.2740–2745.

Arora, S. et al., 2011. Framework for incorporating simulation into urology training. *BJU International*, 107(5), pp.806–810.

Asch, D.A., 2009. Evaluating Obstetrical Residency Programs Using Patient Outcomes. *JAMA*, 302(12), pp.1277–1283.

Asch, D.A. et al., 2014. How Do You Deliver a Good Obstetrician? Outcome-Based Evaluation of Medical Education. *Academic Medicine*, 89(1), pp.24–26.

Association, A.E.R. et al., 1999. *Standards for educational and psychological testing*, Amer Educational Research Assn.

Aydin, A. et al., 2016. Simulation-based training and assessment in urological surgery. *Nature Reviews Urology*, 13(9), pp.503–519.

Azari, D.P. et al., 2017. Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of Surgery*, p.1.

Azhar, R.A. et al., 2016. Enhanced Recovery after Urological Surgery: A Contemporary Systematic Review of Outcomes, Key Elements, and Research Needs. *European Urology*, 70(1), pp.176–187.

Balasundaram, I., Aggarwal, R. & Darzi, A., 2008. Short-phase training on a virtual reality simulator improves technical performance in tele-robotic surgery. *The international journal of medical robotics + computer assisted surgery : MRCAS*, 4(2), pp.139–145.

Ban, K.A. et al., 2016. Evaluation of the ProPublica Surgeon Scorecard “Adjusted Complication Rate” Measure Specifications. *Annals of Surgery*, p.1.

Baron, R.B., 2013. Can we achieve public accountability for graduate medical education outcomes? *Academic medicine : journal of the Association of American Medical Colleges*, 88(9), pp.1199–1201.

Barsuk, J.H. et al., 2009. Mastery Learning of Temporary Hemodialysis Catheter Insertion by Nephrology Fellows Using Simulation Technology and Deliberate Practice. *American Journal of Kidney Diseases*, 54(1), pp.70–76.

Barsuk, J.H., Cohen, E.R., Caprio, T., et al., 2012. Simulation-based education with mastery learning improves residents' lumbar puncture skills. *Neurology*, 79(2), pp.132–137.

Barsuk, J.H., Cohen, E.R., Vozenilek, J.A., et al., 2012. Simulation-Based Education with Mastery Learning Improves Paracentesis Skills. *Journal of Graduate Medical Education*, 4(1), pp.23–27.

Batalden, P. et al., 2002. General Competencies And Accreditation In Graduate Medical Education. *Health Affairs*, 21(5), pp.103–111.

Beal, D.J. et al., 2005. An episodic process model of affective influences on performance. *The Journal of applied psychology*, 90(6), pp.1054–1068.

Beard, J.D., 2005. Setting Standards for the Assessment of Operative Competence. *European Journal of Vascular and Endovascular Surgery*, 30(2), pp.215–218.

Beard, J.D. et al., 2011. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Health Technology Assessment*, 15(1), pp.i–xxi– 1–162.

Bell, R.H. et al., 2009. Operative experience of residents in US general surgery programs: a gap between expectation and experience. *Annals of Surgery*, 249(5), pp.719–724.

Ben-Zvi, T. et al., 2014. Urological resident exposure to transurethral surgical options for BPH management in 2012-2013: A pan-Canadian survey. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 8(1-2), pp.54–60.

Bernard, J.A. et al., 2016. Reliability and Validity of 3 Methods of Assessing Orthopedic Resident Skill in Shoulder Surgery. *Journal of surgical education*, 73(6), pp.1020–1025.

Bhindi, B. et al., 2014. The Importance of Surgeon Characteristics on Impacting Oncologic Outcomes for Patients Undergoing Radical Cystectomy. *The Journal of Urology*, 192(3), pp.714–720.

Bilimoria, K.Y. et al., 2016. National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training. *The New England journal of medicine*, 374(8), pp.713–727.

Birkmeyer, J.D. et al., 2013. Surgical Skill and Complication Rates after Bariatric Surgery. *The New England journal of medicine*, 369(15), pp.1434–1442.

Bleakley, A., 2006. Broadening conceptions of learning in medical education: the message from teamworking. *Medical education*, 40(2), pp.150–157.

Boer, den, M.C. et al., 2018. Ethical dilemmas of recording and reviewing neonatal resuscitation. *Archives of disease in childhood. Fetal and neonatal edition*, pp.fetalneonatal–2017–314191.

Bonrath, E.M. & Gordon, L.E., 2015. Characterising “near miss” events in complex laparoscopic surgery through video analysis. *BMJ Quality & Safety*.

Bonrath, E.M., Dedy, N.J., et al., 2015. Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial. *Annals of Surgery*, 262(2), pp.205–212.

Bonrath, E.M., Dedy, N.J., Zevin, B. & Grantcharov, T.P., 2013a. Defining technical errors in laparoscopic surgery: a systematic review. *Surgical Endoscopy*, 27(8), pp.2678–2691.

Bonrath, E.M., Dedy, N.J., Zevin, B. & Grantcharov, T.P., 2013b. International consensus on safe techniques and error definitions in laparoscopic surgery. *Surgical Endoscopy*, 28(5), pp.1535–1544.

Bonrath, E.M., Gordon, L.E. & Grantcharov, T.P., 2015. Characterising “near miss” events in complex laparoscopic surgery through video analysis. *BMJ Quality & Safety*, 24(8), pp.516–521.

Bonrath, E.M., Zevin, B., et al., 2013. Error rating tool to identify and analyse technical errors and events in laparoscopic surgery. *British Journal of Surgery*, 100(8), pp.1080–1088.

Bookless, L.R., Jones, A.E. & Phillips, A.W., 2015. What evidence is there for the use of workplace-based assessment in surgical training? *Journal of surgical education*, 72(3), pp.367–368.

Brown, K., Mosley, N. & Tierney, J., 2017. Battle of the bots: a comparison of the standard da Vinci and the da Vinci Surgical Skills Simulator in surgical skills acquisition. *Journal of Robotic Surgery*, 11(2), pp.159–162.

Brunckhorst, O. et al., 2015. The Relationship Between Technical And Nontechnical Skills Within A Simulation-Based Ureteroscopy Training Environment. *Journal of surgical education*, 72(5), pp.1039–1044.

Brydges, R. et al., 2015. Linking Simulation-Based Educational Assessments and Patient-Related Outcomes. *Academic Medicine*, 90(2), pp.246–256.

Burch, V.C. et al., 2005. A structured assessment of newly qualified medical graduates. *Medical education*, 39(7), pp.723–731.

Canada, R.C.O.P.A.S.O., 2018. CBD Implementation

. *Royal College of Physicians and Surgeons of Canada*. Available at: http://www.royalcollege.ca/rcsite/cbd/cbd-implementation-e [Accessed April 20, 2018].

Canada, R.C.O.P.A.S.O., 2014. Competence by Design: Reshaping Canadian Medical Education. pp.1–141.

Carraccio, C. et al., 2002. Shifting paradigms: from Flexner to competencies. *Academic Medicine*, 77(5), pp.361–367.

Casalino, L.P. et al., 2016. US Physician Practices Spend More Than $15.4 Billion Annually To Report Quality Measures. *Health affairs (Project Hope)*, 35(3), pp.401–406.

Cate, Ten, O., 2005. Entrustability of professional activities and competency-based training. *Medical education*, 39(12), pp.1176–1177.

Cate, Ten, O., 2013. Nuts and bolts of entrustable professional activities. *Journal of Graduate Medical Education*, 5(1), pp.157–158.

Cate, Ten, O. et al., 2015. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Medical Teacher*, 37(11), pp.983–1002.

Cate, Ten, O. et al., 2016. Entrustment Decision Making in Clinical Training. *Academic Medicine*, 91(2), pp.191–198.

Cate, Ten, O., Snell, L. & Carraccio, C., 2010. Medical competence: the interplay between individual ability and the health care environment. *Medical Teacher*, 32(8), pp.669–675.

Caverzagie, K.J. et al., 2017. Overarching challenges to the implementation of competency-based medical education. *Medical Teacher*, 39(6), pp.588–593.

Cendan, J., Wier, D. & Behrns, K., 2013. A primer on standards setting as it applies to surgical education and credentialing. *Surgical Endoscopy*, 27(7), pp.2631–2637.

Chandra, V. et al., 2010. A comparison of laparoscopic and robotic assisted suturing performance by experts and novices. *Surgery*, 147(6), pp.830–839.

Chowriappa, A. et al., 2015. Augmented-reality-based skills training for robot-assisted urethrovesical anastomosis: a multi-institutional randomised controlled trial. *BJU International*, 115(2), pp.336–345.

Chowriappa, A.J. et al., 2013. Development and validation of a composite scoring system for robot-assisted surgical training--the Robotic Skills Assessment Score. *The Journal of surgical research*, 185(2), pp.561–569.

Christ, T., Arya, P. & Chiu, M.M., 2017. Video use in teacher education: An international survey of practices. *Teaching and Teacher Education*, 63, pp.22–35.

Chung, R.S. & Ahmed, Naveed, 2010. The impact of minimally invasive surgery on residents' open operative experience: analysis of two decades of national data. *Annals of Surgery*, 251(2), pp.205–212.

Cizek, G.J., 1996. Standard-setting guidelines. *Educational Measurement*.

Cohen, E.R. et al., 2013. Raising the Bar: Reassessing Standards for Procedural Competence. *Teaching and learning in medicine*, 25(1), pp.6–9.

Connolly, A., La Cruz, De, J. & Sullivan, S., 2015. A Tool for “Real-Time” Formative and Summative Assessments of Milestones and Surgical Skills. *Obstetrics & Gynecology*, 126, p.51S.

Cook, D.A. & Beckman, T.J., 2006. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, 119(2), pp.166.e7–166.e16.

Cook, D.A. & Hatala, R., 2016. Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation*, 1(1), pp.1–12.

Cook, D.A. & Reed, D.A., 2015. Appraising the Quality of Medical Education Research Methods. *Academic Medicine*, 90(8), pp.1067–1076.

Cook, D.A. et al., 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical education*, 49(6), pp.560–575.

Cook, D.A. et al., 2014. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in health sciences education : theory and practice*, 19(2), pp.233–250.

Cooperberg, M.R., Odisho, A.Y. & Carroll, P.R., 2012. Outcomes for radical prostatectomy: is it the singer, the song, or both? *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 30(5), pp.476–478.

CRONBACH, L.J. & MEEHL, P.E., 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4), pp.281–302.

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. *Medical education*, 46(1), pp.28–37.

Crossley, J. et al., 2011. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Medical education*, 45(6), pp.560–569.

Cui, H.W., Turney, B.W. & Griffiths, J., 2017. The Preoperative Assessment and Optimization of Patients Undergoing Major Urological Surgery. *Current Urology Reports*, 18(7), p.54.

Dagnaes-Hansen, J. et al., 2017. Direct Observation vs. Video-Based Assessment in Flexible Cystoscopy. *Journal of surgical education*, 75(3), pp.671–677.

David Hodges, B., 2010. A Tea-Steeping or i-Doc Model for Medical Education? *Academic Medicine*, 85, pp.S34–S44.

Davies, R.M., Hadfield-Law, L. & Turner, P.G., 2018. Development and Evaluation of a New Formative Assessment of Surgical Performance. *Journal of surgical education*.

Davis, J.W. et al., 2010. Initial experience of teaching robot-assisted radical prostatectomy to surgeons-in-training: can training be evaluated and standardized? *BJU International*, 105(8), pp.1148–1154.

de Montbrun, S. et al., 2016. Implementing and Evaluating a National Certification Technical Skills Examination. *Annals of Surgery*, pp.1–6.

de Montbrun, S., Satterthwaite, L. & Grantcharov, T.P., 2015. Setting pass scores for assessment of technical performance by surgical trainees. *British Journal of Surgery*, 103(3), pp.300–306.

Deadrick, D., 1997. Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management*, 23(6), pp.745–757.

Dedy, N.J. et al., 2015. Implementation of an Effective Strategy for Teaching Nontechnical Skills in the Operating Room: A Single-blinded Nonrandomized Trial. *Annals of Surgery*, p.1.

DeVellis, R.F., 2016. *Scale Development*, SAGE Publications.

Dimick, J.B. & Greenberg, C.C., 2013. Understanding Gaps in Surgical Quality. *Annals of Surgery*, 257(1), pp.6–7.

Dimick, J.B. & Varban, O.A., 2015. Surgical video analysis: an emerging tool for improving surgeon performance. *BMJ Quality & Safety*.

Diwadkar, G.B. et al., 2009. Assessing vaginal surgical skills using video motion analysis. *Obstetrics & Gynecology*, 114(2 Pt 1), pp.244–251.

Dougherty, P. et al., 2013. Intraoperative Assessment of Residents. *Journal of Graduate Medical Education*, 5(2), pp.333–334.

Dougherty, P.J., 2013. What the ACGME's next accreditation system means to you. *Clinical Orthopaedics and Related Research®*, 471(9), pp.2746–2750.

Downing, S.M. & Haladyna, T.M., 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical education*, 38(3), pp.327–333.

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*, 18(1), pp.50–57.

Downing, S.M., Tekian, A. & Yudkowsky, R., 2010. RESEARCH METHODOLOGY: Procedures for Establishing Defensible Absolute Passing Scores on Performance Examinations in Health Professions Education. *Teaching and learning in medicine*, 18(1), pp.50–57.

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

Drolet, B.C. et al., 2013. Surgical residents' perceptions of 2011 Accreditation Council for Graduate Medical Education duty hour regulations. *JAMA Surgery*, 148(5), pp.427–433.

Dubin, A.K. et al., 2017. A Comparison of Robotic Simulation Performance on Basic Virtual Reality Skills: Simulator Subjective Versus Objective Assessment Tools. *Journal of Minimally Invasive Gynecology*.

Dulan, G. et al., 2012. Proficiency-based training for robotic surgery: construct validity, workload, and expert levels for nine inanimate exercises. *Surgical Endoscopy And Other Interventional Techniques*, 26(6), pp.1516–1521.

Eardley, I. et al., 2013. Workplace-based assessment in surgical training: experiences from the Intercollegiate Surgical Curriculum Programme. *ANZ Journal of Surgery*, 83(6), pp.448–453.

Eckert, M. et al., 2010. The changing face of the general surgeon: national and local trends in resident operative experience. *American journal of surgery*, 199(5), pp.652–656.

education, S.D.M.2003, Validity: on the meaningful interpretation of assessment data. *Wiley Online Library*

.

Ellaway, R. et al., 2007. Cross-referencing the Scottish Doctor and Tomorrow's Doctors learning outcome frameworks. *Medical Teacher*, 29(7), pp.630–635.

Elsey, E.J. et al., 2017. Meta-analysis of operative experiences of general surgery trainees during training. *British Journal of Surgery*, 104(1), pp.22–33.

England, T.R.C.O.S.O., 2014. *Good Surgical Practice*,

Englander, R. et al., 2016. Toward Defining the Foundation of the MD Degree: Core Entrustable Professional Activities for Entering Residency. *Academic medicine : journal of the Association of American Medical Colleges*, 91(10), pp.1352–1358.

Epstein, R.M., 2007. Assessment in medical education. *The New England journal of medicine*, 356(4), pp.387–396.

Ericsson, K.A., 2004. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, 79(10 Suppl), pp.S70–81.

Eubanks, T.R. et al., 1999. An objective scoring system for laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 189(6), pp.566–574.

Eva, K.W. et al., 2016. Towards a program of assessment for health professionals: from training into practice. *Advances in health sciences education : theory and practice*, 21(4), pp.897–913.

FACS, T.S.L.M. et al., 2013. Virtual Reality Robotic Surgery Warm-Up Improves Task Performance in a Dry Laboratory Environment: A Prospective Randomized Controlled Study. *Journal of the American College of Surgeons*, 216(6), pp.1181–1192.

Fecso, A.B. et al., 2017. Technical Performance as a Predictor of Complications in Laparoscopic Gastric Cancer Surgery: Another Piece of the Puzzle. *Journal of the American College of Surgeons*, 225(4), p.S92.

Fecso, A.B. et al., 2016. The Effect of Technical Performance on Patient Outcomes in Surgery: A Systematic Review. *Annals of Surgery*.

Fecso, A.B., Bhatti, J.A., et al., 2018. Technical Performance as a Predictor of Clinical Outcomes in Laparoscopic Gastric Cancer Surgery. *Annals of Surgery*, p.1.

Fecso, A.B., Kuzulugil, S.S., et al., 2018. Relationship between intraoperative non-technical performance and technical events in bariatric surgery. *British Journal of Surgery*, 83, p.249−257.

Ferroli, P. et al., 2012. Application of an aviation model of incident reporting and investigation to the neurosurgical scenario: method and preliminary data. *Neurosurgical focus*, 33(5), p.E7.

Finnegan, K.T. et al., 2012. da Vinci Skills Simulator Construct Validation Study: Correlation of Prior Robotic Experience With Overall Score and Time Score Simulator Performance. *Urology*, 80(2), pp.330–336.

Fisher, C.D., 2015. What If We Took Within-Person Performance Variability Seriously? *Industrial and Organizational Psychology*, 1(02), pp.185–189.

Fitzgerald, E. & Giddings, C., 2015. Regional variations in surgical examination performance across the UK: is there a postcode lottery in training? *The Bulletin of the Royal College of Surgeons of England*, 93(6), pp.214–216.

Foell, K., Finelli, A., et al., 2013. Robotic surgery basic skills training: Evaluation of a pilot multidisciplinary simulation-based curriculum. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 7(11-12), pp.430–434.

Foell, K., Furse, A., et al., 2013. Multidisciplinary validation study of the da Vinci Skills Simulator: educational tool and assessment device. *Journal of Robotic Surgery*, 7(4), pp.365–369.

Foster, S.L. & Cone, J.D., 1995. Validity issues in clinical assessment. *Psychological Assessment*, 7(3), pp.248–260.

FRANK, J.E., 2005. The CanMEDS 2005 Physician Competency Framework. *rcpsc.medical.org*.

Frank, J.R. & Danoff, D., 2007. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Medical Teacher*, 29(7), pp.642–647.

Frank, J.R. et al., 2010. Competency-based medical education: theory to practice. *Medical Teacher*, 32(8), pp.638–645.

Frank, J.R. et al., 2017. Implementing competency-based medical education: Moving forward. *Medical Teacher*, 39(6), pp.568–573.

Frank, J.R. et al., 1996. Skills for the new millennium: report of the societal needs working group, CanMEDS 2000 Project. *Ann R Coll Physicians Surg Can*, 29(4), pp.206–216.

Frank, J.R., Snell, L. & Sherbino, J., 2015. *Canmeds 2015 Physician Competency Framework*, Royal College of Physicians and Surgeons of CA.

Fraser, S.A. et al., 2003. Evaluating laparoscopic skills. *Surgical Endoscopy*, 17(6), pp.964–967.

Friad, G., Sabah, K. & Ameen, I.H., 2014. Urology training in the developing world: The trainees' perspective in Kurdistan, Iraq. *Arab Journal of Urology*, 12(1), pp.6–11.

Fried, G.M. et al., 2004. Proving the value of simulation in laparoscopic surgery. *Annals of Surgery*, 240(3), pp.518–25– discussion 525–8.

Friedell, M.L. et al., 2014. Perceptions of graduating general surgery chief residents: are they confident in their training? *Journal of the American College of Surgeons*, 218(4), pp.695–703.

Furriel, F.T.G. et al., 2013. Training of European urology residents in laparoscopy: results of a pan-European survey. *BJU International*, 112(8), pp.1223–1228.

Gavazzi, A. et al., 2011. Face, content and construct validity of a virtual reality simulator for robotic surgery (SEP Robot). *The Annals of The Royal College of Surgeons of England*, 93(2), pp.152–156.

George, B.C. et al., 2017. Readiness of US General Surgery Residents for Independent Practice. *Annals of Surgery*, 266(4), pp.582–594.

Ghani, K.R. et al., 2016. Measuring to Improve: Peer and Crowd-sourced Assessments of Technical Skill with Robot-assisted Radical Prostatectomy. *European Urology*, 69(4), pp.547–550.

Ghayda, R.A. et al., 2017. Andrology/male infertlity subspecialty exposure during U.S based urology residency training. *Fertility and Sterility*, 108(3), pp.e80–e81.

Gillman, L.M. & Vergis, A., 2013. General surgery graduates may be ill prepared to enter rural or community surgical practice. *American journal of surgery*, 205(6), pp.752–757.

Gjeraa, K. et al., 2016. Non-technical skills in minimally invasive surgery teams: a systematic review. *Surgical Endoscopy And Other Interventional Techniques*, 30(12), pp.5185–5199.

Gofton, W.D.N.B.G. & Bhanji, F., 2017. Workplace-Based Assessment Implementation Guide: Formative tips for medical teaching practice. *The Royal College of Physicians and Surgeons of Canada*, pp.1–12. Available at: http://www.royalcollege.ca/rcsite/documents/cbd/wba-implementation-guide-tips-medical-teaching-practice-e.pdf.

Gofton, W.T. et al., 2012. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE). *Academic Medicine*, 87(10), pp.1401–1407.

Goh, A.C. et al., 2012. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of Urology*, 187(1), pp.247–252.

Goh, A.C. et al., 2015. Multi-Institutional Validation of Fundamental Inanimate Robotic Skills Tasks. *The Journal of Urology*, 194(6), pp.1751–1756.

Goldenberg, M.G. & Grantcharov, T.P., 2017. A Novel Method of Setting Performance Standards in Surgery Using Patient Outcomes. *Annals of Surgery*, p.1.

Goldenberg, M.G. et al., 2018. Implementing Assessments of Robotic-Assisted Technical Skill in Urologic Education: A Systematic Review and Synthesis of the Validity Evidence. *BJU International*.

Goldenberg, M.G., Garbens, A., et al., 2017. Systematic review to establish absolute standards for technical performance in surgery. *British Journal of Surgery*, 104(1), pp.13–21.

Goldenberg, M.G., Goldenberg, L. & Grantcharov, T.P., 2017. Surgeon Performance Predicts Early Continence After Robot-Assisted Radical Prostatectomy. *Journal of endourology / Endourological Society*, 31(9), pp.858–863.

Goldenberg, M.G., Jung, J. & Grantcharov, T.P., 2017. Using Data to Enhance Performance and Improve Quality and Safety in Surgery. *JAMA Surgery*.

Gomez, E.D. et al., 2015. Objective assessment of robotic surgical skill using instrument contact vibrations. *Surgical Endoscopy And Other Interventional Techniques*, pp.1–13.

Gostlow, H. et al., 2017. Non-technical skills of surgical trainees and experienced surgeons. *British Journal of Surgery*, 104(6), pp.777–785.

Govaerts, M. & Van Der Vleuten, C.P.M., 2013. Validity in work-based assessment: expanding our horizons. *Medical education*, 47(12), pp.1164–1174.

Govaerts, M.J.B. et al., 2013. Workplace-based assessment: raters' performance theories and constructs. *Advances in health sciences education : theory and practice*, 18(3), pp.375–396.

Great Britain, G.M.C. & Staff, G.M.C.G.B., 2013. *Good Medical Practice*,

Green, I.C. et al., 2013. Creating a Validated Robotic Curriculum for Resident and Fellow Education. *The Journal of Minimally Invasive Gynecology*, 20(S), p.S130.

Green, M.L. et al., 2009. Charting the Road to Competence: Developmental Milestones for Internal Medicine Residency Training. *Journal of Graduate Medical Education*, 1(1), pp.5–20.

Greenberg, C.C. et al., 2017. A Statewide Surgical Coaching Program Provides Opportunity for Continuous Professional Development. *Annals of Surgery*, Publish Ahead of Print, p.1.

Greenberg, C.C. et al., 2015. Surgical coaching for individual performance improvement. *Annals of Surgery*, 261(1), pp.32–34.

Greenberg, C.C., Dombrowski, J. & Dimick, J.B., 2016. Video-Based Surgical Coaching: An Emerging Approach to Performance Improvement. *JAMA Surgery*, 151(3), pp.282–283.

Grillo, H.C., 2004. *Edward D. Churchill and the "rectangular" surgical residency*,

Grober, E.D., Elterman, D.S. & Jewett, M.A.S., 2008. Fellow or foe: the impact of fellowship training programs on the education of Canadian urology residents. *Canadian Urological Association Journal*, 2(1), pp.33–37.

Gurgacz, S.L. et al., 2012. Credentialing of surgeons: a systematic review across a number of jurisdictions. *ANZ Journal of Surgery*, 82(7-8), pp.492–498.

Guru, K.A. et al., 2009. *IN-VIVO VIDEOS ENHANCE COGNITIVE SKILLS FOR DA VINCI® SURGICAL SYSTEM*, The Journal of ….

Habuchi, T. et al., 2011. Evaluation of 2,590 urological laparoscopic surgeries undertaken by urological surgeons accredited by an endoscopic surgical skill qualification system in urological laparoscopy in Japan. *Surgical Endoscopy*, 26(6), pp.1656–1663.

Hakimi, A.A. et al., 2009. Direct Comparison of Surgical and Functional Outcomes of Robotic-Assisted Versus Pure Laparoscopic Radical Prostatectomy: Single-Surgeon Experience. *Urology*, 73(1), pp.119–123.

Halsted, W.S., 1904. *The Training of the surgeon*,

Hammond, L., Ketchum, J. & Schwartz, B.F., 2005. Accreditation Council on Graduate Medical Education Technical Skills Competency Compliance: Urologic Surgical Skills. *Journal of the American College of Surgeons*, 201(3), pp.454–457.

Hanks, J.B. et al., 2011. Feast or famine? The variable impact of coexisting fellowships on general surgery resident operative volumes. *Annals of Surgery*, 254(3), pp.476–83– discussion 483–5.

Harris, P. et al., 2017. Evolving concepts of assessment in a competency-based world. *Medical Teacher*, 39(6), pp.603–608.

Hassan, S.O. et al., 2015. Conventional Laparoscopic vs Robotic Training: Which is Better for Naive Users? A Randomized Prospective Crossover Study. *Journal of surgical education*, 72(4), pp.592–599.

Hatala, R. et al., 2015. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Advances in Health Sciences Education*, 20(5), pp.1–27.

Hauer, K.E. et al., 2018. Translating Theory Into Practice: Implementing a Program of Assessment. *Academic medicine : journal of the Association of American Medical Colleges*, 93(3), pp.444–450.

Hawkins, R.E. et al., 2015. Implementation of competency-based medical education: are we addressing the concerns and challenges? *Medical education*, 49(11), pp.1086–1102.

Haynes, A.B. et al., 2009. A surgical safety checklist to reduce morbidity and mortality in a global population. *The New England journal of medicine*, 360(5), pp.491–499.

Health Quality Ontario, 2017. Robotic Surgical System for Radical Prostatectomy: A Health Technology Assessment. *Ontario health technology assessment series*, 17(11), pp.1–172.

Henken, K.R. et al., 2012. Implications of the law on video recording in clinical practice. *Surgical Endoscopy And Other Interventional Techniques*, 26(10), pp.2909–2916.

Hinata, N. et al., 2013. Dry box training with three-dimensional vision for the assistant surgeon in robot-assisted urological surgery. *International journal of urology : official journal of the Japanese Urological Association*, 20(10), pp.1037–1041.

Hogg, M.E. et al., 2016. Grading of Surgeon Technical Performance Predicts Postoperative Pancreatic Fistula for Pancreaticoduodenectomy Independent of Patient-related Variables. *Annals of Surgery*, 264(3), pp.482–491.

Holmboe, E.S. et al., 2010. The role of assessment in competency-based medical education. *Medical Teacher*, 32(8), pp.676–682.

Holst, D., Kowalewski, T.M., White, L.W., Brand, T.C., Harper, J.D., Sorensen, M.D., et al., 2015. Crowd-Sourced Assessment of Technical Skills (C-SATS): Differentiating Animate Surgical Skill Through the Wisdom of Crowds. *Journal of Endourology*, pp.150413093359007–6.

Holst, D., Kowalewski, T.M., White, L.W., Brand, T.C., Harper, J.D., Sorenson, M.D., et al., 2015. Crowd-Sourced Assessment of Technical Skills: An Adjunct to Urology Resident Surgical Simulation Training. *Journal of Endourology*, 29(5), pp.604–609.

Hopmans, C.J. et al., 2015. Impact of the European Working Time Directive (EWTD) on the operative experience of surgery residents. *Surgery*, 157(4), pp.634–641.

Houston, W.R., 2016. Designing Competency-based Instructional Systems. *Journal of Teacher Education*, 24(3), pp.200–204.

Hu, Y.-Y. et al., 2016. Complementing Operating Room Teaching With Video-Based Coaching. *JAMA Surgery*, 152(4), pp.318–325.

Hu, Y.-Y. et al., 2012. Protecting patients from an unsafe system: the etiology and recovery of intraoperative deviations in care. *Annals of Surgery*, 256(2), pp.203–210.

Huang, G.C. et al., 2009. Procedural competence in internal medicine residents: validity of a central venous catheter insertion assessment instrument. *Academic medicine : journal of the Association of American Medical Colleges*, 84(8), pp.1127–1134.

Hung, A.J. et al., 2013. Comparative assessment of three standardized robotic surgery training methods. *BJU International*, 112(6), pp.864–871.

Hung, A.J. et al., 2012. Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. *The Journal of Urology*, 187(2), pp.630–637.

Hung, A.J. et al., 2015. Development and Validation of a Novel Robotic Procedure Specific Simulation Platform: Partial Nephrectomy. *The Journal of Urology*, 194(2), pp.520–526.

Hung, A.J. et al., 2017. Structured learning for robotic surgery utilizing a proficiency score: a pilot study. *World Journal of Urology*, 35(1), pp.27–34.

Hussein, A.A. et al., 2016. Development and Validation of an Objective Scoring Tool for Robot-Assisted Radical Prostatectomy: Prostatectomy Assessment and Competency Evaluation. *The Journal of Urology*.

Husslein, H. et al., 2015. The Generic Error Rating Tool: A Novel Approach to Assessment of Performance and Surgical Education in Gynecologic Laparoscopy. *Journal of surgical education*, 72(6), pp.1259–1265.

Hwang, H., 2009. Does general surgery residency prepare surgeons for community practice in British Columbia? *Canadian journal of surgery. Journal canadien de chirurgie*, 52(3), pp.196–200.

Ibrahim, A.M., Varban, O.A. & Dimick, J.B., 2016. Novel Uses of Video to Accelerate the Surgical Learning Curve. *Journal of laparoendoscopic & advanced surgical techniques. Part A*, 26(4), pp.240–242.

Iglehart, J.K., 2008. Revisiting duty-hour limits--IOM recommendations for patient safety and resident education. *The New England journal of medicine*, 359(25), pp.2633–2635.

Ilgen, J.S. et al., 2015. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Medical education*, 49(2), pp.161–173.

Imrie, K. et al., *Towards a Pan-canadian Consensus on Resident Duty Hours: Final Report and Recommendations*,

Jacobsen, M.E. et al., 2015. Testing Basic Competency in Knee Arthroscopy Using a Virtual Reality Simulator: Exploring Validity and Reliability. *The Journal of Bone & Joint Surgery*, 97(9), pp.775–781.

Jaeger, R.M., 1989. Selection of Judges for Standard Setting: What Kinds? How Many?.

Jeffcott, S.A., Ibrahim, J.E. & Cameron, P.A., 2009. Resilience in healthcare and clinical handover. *Quality & safety in health care*, 18(4), pp.256–260.

Jelovsek, J.E. et al., 2010. Establishing cutoff scores on assessments of surgical skills to determine surgical competence. *YMOB*, 203(1), pp.81.e1–81.e6.

Jensen, A.R. et al., 2012. Educational feedback in the operating room: a gap between resident and faculty perceptions. *American journal of surgery*, 204(2), pp.248–255.

Joice, P., Hanna, G.B. & Cuschieri, A., 1998. Errors enacted during endoscopic surgery—a human reliability analysis. *Applied ergonomics*, 29(6), pp.409–414.

Jonsson, M.N. et al., 2011. ProMIS ™Can Serve as a da Vinci ®Simulator—A Construct Validity Study. *Journal of Endourology*, 25(2), pp.345–350.

Kairys, J.C. et al., 2008. Cumulative operative experience is decreasing during general surgery residency: a worrisome trend for surgical trainees? *Journal of the American College of Surgeons*, 206(5), pp.804–11– discussion 811–3.

Kane, M.T., 1992. An argument-based approach to validity. *Psychological bulletin*.

Kane, M.T., 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), pp.1–73.

Kane, M.T., Crooks, T.J. & Cohen, A.S., 1999. Designing and Evaluating Standard-Setting Procedures for Licensure and Certification Tests. *Advances in health sciences education : theory and practice*, 4(3), pp.195–207.

Kang, S.G. et al., 2014. The Tube 3 module designed for practicing vesicourethral anastomosis in a virtual reality robotic simulator: determination of face, content, and construct validity. *Urology*, 84(2), pp.345–350.

Kaufman, D.M. et al., 2000. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic medicine : journal of the Association of American Medical Colleges*, 75(3), pp.267–271.

Keith Francis Rourke, A.E.M., 2016. Mapping a competency-based surgical curriculum in urology: Agreement (and discrepancies) in the Canadian national opinion. *Canadian Urological Association Journal*, 10(5-6), pp.161–166.

Kelly, B.D., Curtin, P.D. & Corcoran, M., 2011. The effects of the European Working Time Directive on surgical training: the basic surgical trainee's perspective. *Irish Journal of Medical Science*, 180(2), pp.435–437.

Kelly, D.C. et al., 2012. Face, content, and construct validation of the da Vinci Skills Simulator. *Urology*, 79(5), pp.1068–1072.

Kenney, P.A. et al., 2009. Face, content, and construct validity of dV-trainer, a novel virtual reality simulator for robotic surgery. *Urology*, 73(6), pp.1288–1292.

Kesavadas, T., Kumar, A. & Srimathveeravalli, G., 2009. *Efficacy of Robotic Surgery SImulator (RoSS) for the davinci® surgical system*, The Journal of ….

Kim, J.J. et al., 2012. Independent predictors of recovery of continence 3 months after robot-assisted laparoscopic radical prostatectomy. *Journal of endourology / Endourological Society*, 26(10), pp.1290–1295.

Kim, J.Y. et al., 2015. Concurrent and predictive validation of robotic simulator Tube 3 module. *Korean journal of urology*, 56(11), pp.756–761.

Kim, S.C. et al., 2011. Factors Determining Functional Outcomes After Radical Prostatectomy: Robot-Assisted Versus Retropubic. *European Urology*, 60(3), pp.413–419.

King, C.R. et al., 2015. Development and Validation of a Laparoscopic Simulation Model for Suturing the Vaginal Cuff. *Obstetrics & Gynecology*, 126, pp.27S–35S.

Kissin, E.Y. et al., 2013. Musculoskeletal Ultrasound Training and Competency Assessment Program for Rheumatology Fellows. *Journal of Ultrasound in Medicine*, 32(10), pp.1735–1743.

Kjellsson, G., Clarke, P. & Gerdtham, U.-G., 2014. Forgetting to remember or remembering to forget: A study of the recall period length in health care survey questions. *Journal of Health Economics*, 35, pp.34–46.

Kogan, J.R. & Holmboe, E., 2013. Realizing the promise and importance of performance-based assessment. *Teaching and learning in medicine*, 25 Suppl 1(sup1), pp.S68–74.

Kogan, J.R. et al., 2011. Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical education*, 45(10), pp.1048–1060.

Kogan, J.R., Holmboe, E.S. & Hauer, K.E., 2009. Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA*, 302(12), pp.1316–1326.

Kohlwes, R.J. et al., 2011. Developing Educators, Investigators, and Leaders During Internal Medicine Residency: The Area of Distinction Program. *Journal of Graduate Medical Education*, 3(4), pp.535–540.

Konge, L. et al., 2012. Establishing Pass/Fail Criteria for Bronchoscopy Performance. *Respiration*, 83(2), pp.140–146.

Konge, L. et al., 2013. Using Virtual-Reality Simulation to Assess Performance in Endobronchial Ultrasound. *Respiration*, 86(1), pp.59–65.

Korets, R. et al., 2011. Validating the use of the Mimic dV-trainer for robotic surgery skill acquisition among urology residents. *Urology*, 78(6), pp.1326–1330.

Korndorffer, J.R., Kasten, S.J. & Downing, S.M., 2010. A call for the utilization of consensus standards in the surgical education literature. *American journal of surgery*, 199(1), pp.99–104.

Kowalewski, T. et al., 2015. HIGH-VOLUME ASSESSMENT OF SURGICAL VIDEOS VIA CROWD-SOURCING: THE BASIC LAPAROSCOPIC UROLOGIC SKILLS (BLUS) INITIATIVE. *JURO*, 193(S), p.e393.

Kramp, K.H. et al., 2015. Validity, reliability and support for implementation of independence-scaled procedural assessment in laparoscopic surgery. *Surgical Endoscopy*, 30(6), pp.2288–2300.

Krause, E.A. & Illich, I., 1977. Medical Nemesis: The Expropriation of Health. *Technology and Culture*, 18(4), p.725.

Kumar, R. et al., 2012. Assessing system operation skills in robotic surgery trainees. *The international journal of medical robotics + computer assisted surgery : MRCAS*, 8(1), pp.118–124.

L LA PIETRA, L.C.L.M.R.Q.S.B., 2006. Medical errors and clinical risk management: state of the art. pp.1–8.

Langerman, A. & Grantcharov, T.P., 2017. Are We Ready for Our Close-up?: Why and How We Must Embrace Video in the OR. *Annals of Surgery*.

Larsen, C.R. et al., 2008. Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale. *BJOG : an international journal of obstetrics and gynaecology*, 115(7), pp.908–916.

Lavigueur-Blouin, H. et al., 2015. Predictors of early continence following robot-assisted radical prostatectomy. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 9(1-2), pp.e93–7.

Lawson, K.A. et al., 2017. Benchmarking quality for renal cancer surgery: Canadian Kidney Cancer information system (CKCis) perspective. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 11(8), pp.232–237.

Lee, G.I. & Lee, M.R., 2017. Can a virtual reality surgical simulation training provide a self-driven and mentor-free skills learning? Investigation of the practical influence of the performance metrics from the virtual reality robotic surgery simulator on the skill learning and associated cognitive workloads. *Surgical Endoscopy And Other Interventional Techniques*, 7(5), pp.431–11.

Lee, J.Y. et al., 2011. Best Practices for Robotic Surgery Training and Credentialing. *The Journal of Urology*, 185(4), pp.1191–1197.

Lee, J.Y., Mucksavage, P., Canales, C., et al., 2012. High Fidelity Simulation Based Team Training in Urology: A Preliminary Interdisciplinary Study of Technical and Nontechnical Skills in Laparoscopic Complications Management. *JURO*, 187(4), pp.1385–1391.

Lee, J.Y., Mucksavage, P., Kerbl, D.C., et al., 2012. Validation study of a virtual reality robotic simulator--role as an assessment tool? *The Journal of Urology*, 187(3), pp.998–1002.

Lee, Y.L., Kilic, G.S. & Phelps, J.Y., 2011. Medicolegal review of liability risks for gynecologists stemming from lack of training in robot-assisted surgery. *Journal of Minimally Invasive Gynecology*, 18(4), pp.512–515.

Legislature, S.O.W., 2015. *An act relating to: video recording of surgical procedures*,

Lendvay, T.S. et al., 2013. Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. *Journal of the American College of Surgeons*, 216(6), pp.1181–1192.

Lendvay, T.S., Casale, P., Sweet, R. & Peters, C., 2008a. Initial validation of a virtual-reality robotic simulator. *Journal of Robotic Surgery*, 2(3), pp.145–149.

Lendvay, T.S., Casale, P., Sweet, R. & Peters, C., 2008b. VR robotic surgery: randomized blinded study of the dV-Trainer robotic simulator. *Studies in health technology and informatics*, 132, pp.242–244.

Lerner, M.A. et al., 2010. Does training on a virtual reality robotic simulator improve performance on the da Vinci surgical system? *Journal of endourology / Endourological Society*, 24(3), pp.467–472.

Liberati, A. et al., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. In BMJ (Clinical research ed.). p. b2700.

Lin, H. et al., 2016. A Narrative Review of High-Quality Literature on the Effects of Resident Duty Hours Reforms. *Academic Medicine*, 91(1), pp.140–150.

Lindenauer, P.K. et al., 2014. Attitudes of hospital leaders toward publicly reported measures of health care quality. *JAMA internal medicine*, 174(12), pp.1904–1911.

Lingard, L., 2005. Getting teams to talk: development and pilot implementation of a checklist to promote interprofessional communication in the OR. *Quality & safety in health care*, 14(5), pp.340–346.

Liss, M.A. et al., 2012. Validation, correlation, and comparison of the da Vinci trainer(™) and the daVinci surgical skills simulator(™) using the Mimic(™) software for urologic robotic surgical education. *Journal of endourology / Endourological Society*, 26(12), pp.1629–1634.

Liss, M.A. et al., 2015. Virtual reality suturing task as an objective test for robotic experience assessment. *BMC Urology*, 15(1), pp.63–7.

Liu, M. et al., 2017. Assessment of Robotic Console Skills (ARCS): construct validity of a novel global rating scale for technical skills in robotically assisted surgery. *Surgical Endoscopy*, 94(5), p.373.

Livingston, S.A. & Zieky, M.J., 1982. *Passing scores*,

Lloyd, D.M., 2011. Robots don't perform surgery, surgeons do. *Bmj*, 343(oct26 1), pp.d6830–d6830.

Lockyer, J. et al., 2017. Core principles of assessment in competency-based medical education. *Medical Teacher*, 39(6), pp.609–616.

Louridas, M. et al., 2017. Practice does not always make perfect: need for selection curricula in modern surgical training. *Surgical Endoscopy And Other Interventional Techniques*, 197, p.447.

Louridas, M., Szasz, P., de Montbrun, S., et al., 2016. Can We Predict Technical Aptitude?: A Systematic Review. *Annals of Surgery*, 263(4), pp.673–691.

Louridas, M., Szasz, P., Montbrun, S. de, et al., 2016. Optimizing the Selection of General Surgery Residents: A National Consensus. *Journal of surgical education*.

Lovegrove, C. et al., 2015. Structured and Modular Training Pathway for Robot-assisted Radical Prostatectomy (RARP): Validation of the RARP Assessment Score and Learning Curve Assessment. *European Urology*.

Lowrance, W.T. & Parekh, D.J., 2012. The rapid uptake of robotic prostatectomy and its collateral effects. *Cancer*, 118(1), pp.4–7.

Lörwald, A.C. et al., 2018. Factors influencing the educational impact of Mini-CEX and DOPS: A qualitative synthesis. *Medical Teacher*, 40(4), pp.414–420.

Lyons, C. et al., 2013. Which skills really matter? proving face, content, and construct validity for a commercial robotic simulator. *Surgical Endoscopy And Other Interventional Techniques*, 27(6), pp.2020–2030.

Maas, M.B., Jaff, M.R. & Rordorf, G.A., 2013. Risk Adjustment for Case Mix and the Effect of Surgeon Volume on Morbidity. *JAMA Surgery*, 148(6), pp.532–5.

Maggard-Gibbons, M., 2014. The use of report cards and outcome measurements to improve the safety of surgical care: the American College of Surgeons National Surgical Quality Improvement Program. *BMJ Quality & Safety*, 23(7), pp.589–599.

Mahmood, N., 2015. CPD for community urologists. *Canadian Urological Association journal = Journal de l'Association des urologues du Canada*, 9(11-12), p.370.

Makary, M.A., 2013. The power of video recording: taking quality to the next level. *JAMA*, 309(15), pp.1591–1592.

Malangoni, M.A. et al., 2013. Operative experience of surgery residents: trends and challenges. *Journal of surgical education*, 70(6), pp.783–788.

Mamut, A.E. et al., 2011. Surgical Case Volume in Canadian Urology Residency: A Comparison of Trends in Open and Minimally Invasive Surgical Experience. *Journal of Endourology*, 25(6), pp.1063–1067.

Mark, J.R. et al., 2014. The effects of fatigue on robotic surgical skill training in Urology residents. *Journal of Robotic Surgery*, 8(3), pp.269–275.

Martin, J.A. et al., 1997. Objective structured assessment of technical skill (OSATS) for surgical residents. *The British journal of surgery*, 84(2), pp.273–278.

Matsuda, T. et al., 2014. Reliability of Laparoscopic Skills Assessment on Video: 8-Year Results of the Endoscopic Surgical Skill Qualification System in Japan. *Journal of Endourology*, 28(11), pp.1374–1378.

Matsuda, T. et al., 2006. The Endoscopic Surgical Skill Qualification System in Urological Laparoscopy: A Novel System in Japan. *The Journal of Urology*, 176(5), pp.2168–2172.

Mattar, S.G. et al., 2013. General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors. *Annals of Surgery*, 258(3), pp.440–449.

Maudsley, W.C., 1996. *Report of the task force to review fundamental issues in specialty education.*

Ottawa: Royal College of Physicians and Surgeons of Canada.

McCluney, A.L. et al., 2007. FLS simulator performance predicts intraoperative laparoscopic skill. *Surgical Endoscopy*, 21(11), pp.1991–1995.

McCoy, A.C. et al., 2013. Are open abdominal procedures a thing of the past? An analysis of graduating general surgery residents' case logs from 2000 to 2011. *Journal of surgical education*, 70(6), pp.683–689.

McCulloch, P. et al., 2009. The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. *Quality & safety in health care*, 18(2), pp.109–115.

McDonough, P.S. et al., 2011. Initial validation of the ProMIS surgical simulator as an objective measure of robotic task performance. *Journal of Robotic Surgery*, 5(3), pp.195–199.

McDougall, E.M., 2007. Validation of Surgical Simulators. *Journal of Endourology*, 21(3), pp.244–247.

McGaghie, W.C.A.O., 1978. Competency-Based Curriculum Development in Medical Education. An Introduction. Public Health Papers No. 68.

McIntyre, H.F. et al., 2010. Implementation of the European Working Time Directive in an NHS trust: impact on patient care and junior doctor welfare. *Clinical medicine (London, England)*, 10(2), pp.134–137.

McVey, R. et al., 2016. Baseline Laparoscopic Skill May Predict Baseline Robotic Skill and Early Robotic Surgery Learning Curve. *Journal of endourology / Endourological Society*, p.end.2015.0774.

MD, E.N.T. et al., 2014. A simulator-based resident curriculum for laparoscopic common bile duct exploration. *Surgery*, 156(4), pp.880–893.

MD, J.R.P.I., 2016. Assessment of Competence. *Surgical Clinics of NA*, 96(1), pp.15–24.

MD, N.N. et al., 2015. Assessment of colonoscopy by use of magnetic endoscopic imaging: design and validation of an automated tool. *Gastrointestinal Endoscopy*, 81(3), pp.548–554.

MD, P.S. et al., 2016. 8th Annual American College of Surgeons Accredited Educational Institutes (ACS-AEI) Consortium Simulation-based summative assessments in surgery. *Surgery*, 160(3), pp.528–535.

Meier, M., Horton, K. & John, H., 2016. Da Vinci© Skills Simulator™: is an early selection of talented console surgeons possible? *Journal of Robotic Surgery*, 10(4), pp.289–296.

Mellinger, J.D., Damewood, R. & Morris, J.B., 2015. Assessing the Quality of Graduate Surgical Training Programs: Perception vs Reality. *Journal of the American College of Surgeons*, 220(5), pp.785–789.

Messick, S., 1975. The standard problem: Meaning and values in measurement and evaluation. *American psychologist*, 30(10), pp.955–966.

Messick, S., 1994. *Validity of Psychological Assessment*,

Meyerson, S.L. et al., 2014. Defining the autonomy gap: when expectations do not meet reality in the operating room. *Journal of surgical education*, 71(6), pp.e64–72.

Miller, G.E., 1990. The assessment of clinical skills/competence/performance. *Academic medicine : journal of the Association of American Medical Colleges*, 65(9 Suppl), pp.S63–7.

Mills, J.T. et al., 2017. Does Robotic Surgical Simulator Performance Correlate With Surgical Skill? *Journal of surgical education*.

Min, H. et al., 2015. Systematic review of coaching to enhance surgeons' operative performance. *Surgery*.

Mishra, A. et al., 2007. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surgical Endoscopy*, 22(1), pp.68–73.

Moglia, A. et al., 2015. A Systematic Review of Virtual Reality Simulators for Robot-assisted Surgery. *European Urology*.

Montorsi, F. et al., 2012. Best practices in robot-assisted radical prostatectomy: recommendations of the Pasadena Consensus Panel. In European urology. pp. 368–381.

Nadler, A. et al., 2015. Career plans and perceptions in readiness to practice of graduating general surgery residents in Canada. *Journal of surgical education*, 72(2), pp.205–211.

Napolitano, L.M. et al., 2014. Are general surgery residents ready to practice? A survey of the American College of Surgeons Board of Governors and Young Fellows Association. *Journal of the American College of Surgeons*, 218(5), pp.1063–1072.e31.

Norcini, J. et al., 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), pp.206–214.

Norcini, J.J., 2005. Current perspectives in assessment: the assessment of performance at work. *Medical education*, 39(9), pp.880–889.

Norcini, J.J., 2003. Setting standards on educational tests. *Medical education*, 37(5), pp.464–469.

Norcini, J.J. et al., 2003. The Mini-CEX: A Method for Assessing Clinical Skills. *Annals of internal medicine*, 138(6), pp.476–481.

Noureldin, Y.A. et al., 2016. Objective Structured Assessment of Technical Skills for the Photoselective Vaporization of the Prostate Procedure: A Pilot Study. *Journal of endourology / Endourological Society*, 30(8), pp.923–929.

Nousiainen, M.T. et al., 2018. Eight-year outcomes of a competency-based residency training program in orthopedic surgery. *Medical Teacher*, 62, pp.1–13.

Nousiainen, M.T. et al., 2017. Implementing competency-based medical education: What changes in curricular structure and processes are needed? *Medical Teacher*, 39(6), pp.594–598.

O'Mahoney, P.R.A. et al., 2016. Driving Surgical Quality Using Operative Video. *Surgical Innovation*, 23(4), pp.337–340.

O'Shea, J.S., 2008. *Becoming a surgeon in the early 20th century: parallels to the present*,

Osman, H. et al., 2015. Are general surgery residents adequately prepared for hepatopancreatobiliary fellowships? A questionnaire-based study. *HPB : the official journal of the International Hepato Pancreato Biliary Association*, 17(3), pp.265–271.

Panzer, R.J. et al., 2013. Increasing Demands for Quality Measurement. *JAMA*, 310(18), pp.1971–1980.

Parker, D.C. et al., 2016. Trends in Urology Residents' Exposure to Operative Urotrauma: A Survey of Residency Program Directors. *URL*, 87, pp.18–24.

Parsa, C.J., Organ, C.H. & Barkan, H., 2000. Changing Patterns of Resident Operative Experience From 1990 to 1997. *Archives of surgery (Chicago, Ill. : 1960)*, 135(5), pp.570–575. Available at: https://www.acgme.org/Portals/0/PFAssets/ProgramRequirements/CPRs_2017-07-01.pdf.

Parsons, B.A. et al., 2011. Surgical training: the impact of changes in curriculum and experience. *Journal of surgical education*, 68(1), pp.44–51.

Paterson, C. et al., 2016. Videotaping of surgical procedures and outcomes following extraperitoneal laparoscopic radical prostatectomy for clinically localized prostate cancer. *Journal of surgical oncology*, 114(8), pp.1016–1023.

Pattani, R., Wu, P.E. & Dhalla, I.A., 2014. Resident duty hours in Canada: past, present and future. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, 186(10), pp.761–765.

Pedersen, P. et al., 2014. Virtual-reality simulation to assess performance in hip fracture surgery. *Acta Orthopaedica*, 85(4), pp.403–407.

Pelgrim, E.A.M. et al., 2011. In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in health sciences education : theory and practice*, 16(1), pp.131–142.

Pellegrini, C.A., 2006. Surgical education in the United States: navigating the white waters., 244(3), pp.335–342.

Pena, G. et al., 2015. Nontechnical skills training for the operating room: A prospective study using simulation and didactic workshop. *Surgery*, 158(1), pp.300–309.

Penson, D.F., 2012. Re: Postgame analysis: using video-based coaching for continuous professional development. *The Journal of Urology*, 188(2), p.561.

Pereira, E.A. & Dean, B.J., 2009. British surgeons' experiences of mandatory online workplace-based assessment. *Journal of the Royal Society of Medicine*, 102(7), pp.287–293.

Perrenot, C. et al., 2012. The virtual reality simulator dV-Trainer(®) is a valid assessment tool for robotic surgical skills. *Surgical Endoscopy And Other Interventional Techniques*, 26(9), pp.2587–2593.

Peters, H. et al., 2017. Twelve tips for the implementation of EPAs for assessment and entrustment decisions. *Medical Teacher*, 39(8), pp.802–807.

Peterson, S.E., 1981. Association of American Medical Colleges. *American Journal of Evaluation*, 2(1), pp.17–19.

Peyré, S.E. et al., 2010. Reliability of a procedural checklist as a high-stakes measurement of advanced technical skill. *The American Journal of Surgery*, 199(1), pp.110–114.

Phé, V. et al., 2017. Outcomes of a virtual-reality simulator-training programme on basic surgical skills in robot-assisted laparoscopic surgery. *The international journal of medical robotics + computer assisted surgery : MRCAS*, 13(2), p.e1740.

Phillips, A.W. et al., 2015. Surgical Trainers' Experience and Perspectives on Workplace-Based Assessments. *JSURG*, 72(5), pp.979–984.

Pollett, W.G. & Dicks, E., 2005. Training of Canadian general surgeons: are they really prepared? CAGS questionnaire on surgical training. *Canadian journal of surgery. Journal canadien de chirurgie*, 48(3), pp.219–224.

Powers, M.K. et al., 2015. Crowdsourcing Assessment of Surgeon Dissection of Renal Artery and Vein During Robotic Partial Nephrectomy: A Novel Approach for Quantitative Assessment of Surgical Performance. *Journal of Endourology*, pp.end.2015.0665–6.

Pradarelli, J.C., Campbell, D.A. & Dimick, J.B., 2015. Hospital credentialing and privileging of surgeons: a potential safety blind spot. *JAMA*, 313(13), pp.1313–1314.

Preisler, L. et al., 2015. Simulation-Based Training for Colonoscopy. *Medicine*, 94(4), pp.e440–8.

Prigoff, J.G., Sherwin, M. & Divino, C.M., 2016. Ethical Recommendations for Video Recording in the Operating Room. *Annals of Surgery*, 264(1), pp.34–35.

Pugh, D. et al., 2017. Using the Entrustable Professional Activities Framework in the Assessment of Procedural Skills. *Journal of Graduate Medical Education*, 9(2), pp.209–214.

Radford, P.D. et al., 2015. Publication of surgeon specific outcome data: A review of implementation, controversies and the potential impact on surgical training. *International Journal of Surgery*, 13, pp.211–216.

Raison, N., Ahmed, K., et al., 2017. Competency based training in robotic surgery: benchmark scores for virtual reality robotic simulation. *BJU International*, 119(5), pp.804–811.

Raison, N., Wood, T., et al., 2017. Development and validation of a tool for non-technical skills evaluation in robotic surgery-the ICARS system. *Surgical Endoscopy And Other Interventional Techniques*, 7(7), pp.403–8.

Ramos, P. et al., 2014. Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. *BJU International*, 113(5), pp.836–842.

Rashid, H.H. et al., 2006. Robotic surgical education: a systematic approach to training urology residents to perform robotic-assisted laparoscopic radical prostatectomy. *Urology*, 68(1), pp.75–79.

Raza, S.J. et al., 2015. Surgical Competency for Urethrovesical Anastomosis During Robot-assisted Radical Prostatectomy: Development and Validation of the Robotic Anastomosis Competency Evaluation. *Urology*, 85(1), pp.27–32.

Reason, J., 1995. Understanding adverse events: human factors. *Quality in health care: QHC*, 4(2), pp.80–89.

Reed, D.A. et al., 2007. Association between funding and quality of published medical education research. *JAMA*, 298(9), pp.1002–1009.

Regehr, G. et al., 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine*, 73(9), p.993.

Rekman, J. et al., 2016. A New Instrument for Assessing Resident Competence in Surgical Clinic: The Ottawa Clinic Assessment Tool. *Journal of surgical education*, 73(4), pp.575–582.

Resnick, M.J. et al., 2013. Long-Term Functional Outcomes after Treatment for Localized Prostate Cancer. *The New England journal of medicine*, 368(5), pp.436–445.

Rex, D.K. et al., 2010. The impact of videorecording on the quality of colonoscopy performance: a pilot study. *The American Journal of Gastroenterology*, 105(11), pp.2312–2317.

Reznick, R. et al., 1997. Testing technical skill via an innovative “bench station” examination. *The American Journal of Surgery*, 173(3), pp.226–230.

Rogers, D.A., Regehr, G. & MacDonald, J., 2002. A role for error training in surgical technical skill instruction and evaluation. *The American Journal of Surgery*, 183(3), pp.242–245.

Ruparel, R.K. et al., 2014. Assessment of virtual reality robotic simulation performance by urology resident trainees. *Journal of surgical education*, 71(3), pp.302–308.

Sachdeva, A.K. & Russell, T.R., 2007. Safe introduction of new procedures and emerging technologies in surgery: education, credentialing, and privileging. *Surgical Oncology Clinics of North America*, 16(1), pp.101–114.

Sandhu, G. et al., 2018. Association of Faculty Entrustment With Resident Autonomy in the Operating Room. *JAMA Surgery*.

Sarker, S.K. et al., 2005. Technical skills errors in laparoscopic cholecystectomy by expert surgeons. *Surgical Endoscopy*, 19(6), pp.832–835.

Sarmento, H. et al., 2014. Match analysis in football: a systematic review. *Journal of sports sciences*, 32(20), pp.1831–1843.

Schiff, L.D., Matthew, P. & Ganesa, W., 2013. Do Current Systems of Credentialing Ensure Patient Safety in Robotic Gynecologic Surgery? A Survey Study. *The Journal of Minimally Invasive Gynecology*, 20(S), pp.S32–S33.

Schommer, E. et al., 2017. Diffusion of Robotic Technology Into Urologic Practice has Led to Improved Resident Physician Robotic Skills. *Journal of surgical education*, 74(1), pp.55–60.

Schuwirth, L.W.T. & Van Der Vleuten, C.P.M., 2011a. Programmatic assessment and Kane's validity perspective. *Medical education*, 46(1), pp.38–48.

Schuwirth, L.W.T. & Van Der Vleuten, C.P.M., 2011b. Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, 33(6), pp.478–485.

Sedlack, R.E., 2011. Training to competency in colonoscopy: assessing and defining competency standards. *Gastrointestinal Endoscopy*, 74(2), pp.355–366.e2.

Sedlack, R.E. & Coyle, W.J., 2016. Assessment of competency in endoscopy: establishing and validating generalizable competency benchmarks for colonoscopy. *Gastrointestinal Endoscopy*, 83(3), pp.516–523.e1.

Seixas-Mikelus, S.A. et al., 2010. Face validation of a novel robotic surgical simulator. *Urology*, 76(2), pp.357–360.

Sethi, A.S. et al., 2009. Validation of a novel virtual reality robotic simulator. *Journal of endourology / Endourological Society*, 23(3), pp.503–508.

Shalhoub, J., Vesey, A.T. & Fitzgerald, J.E.F., 2014. What Evidence is There for the Use of Workplace-Based Assessment in Surgical Training? *Journal of surgical education*, 71(6), pp.906–915.

Shamim Khan, M. et al., 2013. Development and implementation of centralized simulation training: evaluation of feasibility, acceptability and construct validity. *BJU International*, 111(3), pp.518–523.

Sherman, K.L. et al., 2013. Surgeons' perceptions of public reporting of hospital and individual surgeon quality. *Medical care*, 51(12), pp.1069–1075.

Siam, B. et al., 2017. Comparison of Appendectomy Outcomes Between Senior General Surgeons and General Surgery Residents. *JAMA Surgery*, 152(7), pp.679–685.

Siddiqui, N.Y. et al., 2014. Validity and Reliability of the Robotic Objective Structured Assessment of Technical Skills. *Obstetrics & Gynecology*, 123(6), pp.1193–1199.

Singh, P. et al., 2014. A Global Delphi Consensus Study on Defining and Measuring Quality in Surgical Training. *Journal of the American College of Surgeons*, 219(3), pp.346–353.e7.

Snaineh, S.T.A. & Seales, B., 2015. Minimally invasive surgery skills assessment using multiple synchronized sensors. In 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). IEEE, pp. 314–319.

Snell, L.S. & Frank, J.R., 2010. Competencies, the tea bag model, and the end of time. *Medical Teacher*, 32(8), pp.629–630.

Society, A.O.W.S. & Churchill, E.D., *Postdoctoral education and training of the surgeon*,

Song, P.H. & Ko, Y.H., 2016. The Surgical Skill of a Novice Trainee Manifests in Time-Consuming Exercises of a Virtual Simulator Rather Than a Quick-Finishing Counterpart: A Concurrent Validity Study Using an Urethrovesical Anastomosis Model. *Journal of surgical education*, 73(1), pp.166–172.

Soucisse, M.L. et al., 2016. Video Coaching as an Efficient Teaching Method for Surgical Residents-A Randomized Controlled Trial. *Journal of surgical education*, 74(2), pp.365–371.

Specialties, A.B.O.M., 2015. *Standards for the ABMS Program for Maintenance of Certification (MOC)*, American Board of Medical Specialties.

Sroka, G. et al., 2010. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. *American journal of surgery*, 199(1), pp.115–120.

Stefanidis, D. et al., 2006. Psychomotor testing predicts rate of skill acquisition for proficiency-based laparoscopic skills training. *Surgery*, 140(2), pp.252–262.

Stegemann, A.P. et al., 2013. Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum. *Urology*, 81(4), pp.767–774.

Stewart, G.L. & Nandkeolyar, A.K., 2007. Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *The Journal of applied psychology*, 92(4), pp.1149–1158.

Stitzenberg, K.B. et al., 2012. Trends in radical prostatectomy: centralization, robotics, and access to urologic cancer care. *Cancer*, 118(1), pp.54–62.

Sturman, M.C., Cheramie, R.A. & Cashen, L.H., 2005. The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *The Journal of applied psychology*, 90(2), pp.269–283.

Sturmberg, J.P. & Hinchy, J., 2010. Borderline competence - from a complexity perspective: conceptualization and implementation for certifying examinations. *Journal of Evaluation in Clinical Practice*, 16(4), pp.867–872.

Svendsen, M.B., 2014. Using motion capture to assess colonoscopy experience level. *World Journal of Gastrointestinal Endoscopy*, 6(5), pp.193–8.

Swanstrom, L.L. et al., 2006. Beta test results of a new system assessing competence in laparoscopic surgery. *Journal of the American College of Surgeons*, 202(1), pp.62–69.

Swing, S.R., 2009. The ACGME outcome project: retrospective and prospective. *Medical Teacher*, 29(7), pp.648–654.

Szasz, P. et al., 2014. Assessing Technical Competence in Surgical Trainees: A Systematic Review. *Annals of Surgery*, 261(6), pp.1–1055.

Szasz, P. et al., 2016. Setting Performance Standards for Technical and Nontechnical Competence in General Surgery. *Annals of Surgery*, Publish Ahead of Print, p.1.

Szymanski, K.M. et al., 2010. Development and validation of an abbreviated version of the expanded prostate cancer index composite instrument for measuring health-related quality of life among prostate cancer survivors. *Urology*, 76(5), pp.1245–1250.

Tam, V., Zeh, H.J. & Hogg, M.E., 2017. Incorporating Metrics of Surgical Proficiency Into Credentialing and Privileging Pathways. *JAMA Surgery*.

Tanaka, A. et al., 2015. Robotic surgery simulation validity and usability comparative analysis. *Surgical Endoscopy And Other Interventional Techniques*, pp.1–10.

Tang, B. et al., 2006. Competence Assessment of Laparoscopic Operative and Cognitive Skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World Journal of Surgery*, 30(4), pp.527–534.

Tang, B., Hanna, G.B. & Cuschieri, A., 2005. Analysis of errors enacted by surgical trainees during skills training courses. *Surgery*, 138(1), pp.14–20.

Tarr, M.E. et al., 2014. Robotic objective structured assessment of technical skills: a randomized multicenter dry laboratory training pilot study. *Female pelvic medicine & reconstructive surgery*, 20(4), pp.228–236.

Tausch, T.J. et al., 2012. Content and construct validation of a robotic surgery curriculum using an electromagnetic instrument tracker. *The Journal of Urology*, 188(3), pp.919–923.

Teman, N.R. et al., 2014. Entrustment of general surgery residents in the operating room: factors contributing to provision of resident autonomy. *Journal of the American College of Surgeons*, 219(4), pp.778–787.

Thinggaard, E. et al., 2015. Validity of a cross-specialty test in basic laparoscopic techniques (TABLT). *British Journal of Surgery*, 102(9), pp.1106–1113.

Thomsen, A.S.S. et al., 2015. Simulation-based certification for cataract surgery. *Acta Ophthalmologica*, 93(5), pp.416–421.

Tiferes, J. et al., 2016. The Loud Surgeon Behind the Console: Understanding Team Activities During Robot-Assisted Surgery. *Journal of surgical education*, 73(3), pp.504–512.

Tjiam, I.M. et al., 2012. Program for laparoscopic urological skills assessment: Setting certification standards for residents. *Minimally invasive therapy & allied technologies : MITAT : official journal of the Society for Minimally Invasive Therapy*, 22(1), pp.26–32.

Tolsgaard, M.G. et al., 2014. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. *Ultrasound in Obstetrics & Gynecology*, 43(4), pp.437–443.

Toner, J. & Moran, A., 2014. In praise of conscious awareness: a new framework for the investigation of “continuous improvement” in expert athletes. *Frontiers in psychology*, 5, p.769.

Touijer, K. et al., 2005. Quality improvement in laparoscopic radical prostatectomy for pT2 prostate cancer: impact of video documentation review on positive surgical margin. *JURO*, 173(3), pp.765–768.

Training, J.C.O.S., 2017. *Guidelines for ARCP at the end of Core Surgical Training*, Available at: https://www.jcst.org/quality-assurance/certification-guidelines-and-checklists/.

Trinh, Q.-D. et al., 2013. A Systematic Review of the Volume–Outcome Relationship for Radical Prostatectomy. *European Urology*, 64(5), pp.786–798.

Tunitsky, E. et al., 2013. Development and validation of a ureteral anastomosis simulation model for surgical training. *Female pelvic medicine & reconstructive surgery*, 19(6), pp.346–351.

Urbach, D.R. et al., 2014. Introduction of surgical safety checklists in Ontario, Canada. *The New England journal of medicine*, 370(11), pp.1029–1038.

van der Vleuten, C.P., 1996. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*, 1(1), pp.41–67.

Van Der Vleuten, C.P.M. & Schuwirth, L.W.T., 2005. Assessing professional competence: from methods to programmes. *Medical education*, 39(3), pp.309–317.

Vassiliou, M.C. et al., 2005. A global assessment tool for evaluation of intraoperative laparoscopic skills. *The American Journal of Surgery*, 190(1), pp.107–113.

Vassiliou, M.C. et al., 2013. Fundamentals of endoscopic surgery: creation and validation of the hands-on test. *Surgical Endoscopy*, 28(3), pp.704–711.

Verheggen, M.M. et al., 2008. Is an Angoff standard an indication of minimal competence of examinees or of judges? *Advances in health sciences education : theory and practice*, 13(2), pp.203–211.

Vernez, S.L. et al., 2016. C-SATS: Assessing Surgical Skills Among Urology Residency Applicants. *Journal of endourology / Endourological Society*, p.end.2016.0569.

Vernez, S.L. et al., 2017. C-SATS: Assessing Surgical Skills Among Urology Residency Applicants. *Journal of Endourology*, 31(S1), pp.S–95–S–100.

Vickers, A. et al., 2011. Cancer control and functional outcomes after radical prostatectomy as markers of surgical quality: analysis of heterogeneity between surgeons at a single cancer center. *European Urology*, 59(3), pp.317–322.

Vlaovic, P.D. et al., 2008. Immediate impact of an intensive one-week laparoscopy training program on laparoscopic skills among postgraduate urologists. *JSLS : Journal of the Society of Laparoendoscopic Surgeons / Society of Laparoendoscopic Surgeons*, 12(1), pp.1–8.

Volpe, A. et al., 2015. Pilot Validation Study of the European Association of Urology Robotic Training Curriculum. *European Urology*, 68(2), pp.292–299.

Walsh, P.C. et al., 2000. Use of intraoperative video documentation to improve sexual function after radical retropubic prostatectomy. *Urology*, 55(1), pp.62–67.

Walzak, A. et al., 2015. Diagnosing Technical Competence in Six Bedside Procedures. *Academic Medicine*, 90(8), pp.1100–1108.

Wayne, D.B. et al., 2007. Do baseline data influence standard setting for a clinical skills examination? *Academic medicine : journal of the Association of American Medical Colleges*, 82(10 Suppl), pp.S105–8.

Wayne, D.B. et al., 2008. Mastery learning of thoracentesis skills by internal medicine residents using simulation technology and deliberate practice. *Journal of Hospital Medicine*, 3(1), pp.48–54.

Wei, J.T. et al., 2000. Development and validation of the expanded prostate cancer index composite (EPIC) for comprehensive assessment of health-related quality of life in men with prostate cancer. *Urology*, 56(6), pp.899–905.

Wenghofer, E.F., Williams, A.P. & Klass, D.J., 2009. Factors affecting physician performance: implications for performance improvement and governance. *Healthcare policy = Politiques de sante*, 5(2), pp.e141–60.

Werner, R.M. & Asch, D.A., 2005. The unintended consequences of publicly reporting quality information. *JAMA*, 293(10), pp.1239–1244.

White, L.W. et al., 2015. Crowd-Sourced Assessment of Technical Skill: A Valid Method for Discriminating Basic Robotic Surgery Skills. *Journal of endourology / Endourological Society*, 29(11), pp.1295–1301.

Whitehurst, S.V. et al., 2015. Comparison of Two Simulation Systems to Support Robotic-Assisted Surgical Training: A Pilot Study (Swine Model). *Journal of Minimally Invasive Gynecology*, 22(3), pp.483–488.

Whittaker, G. et al., 2016. Validation of the RobotiX Mentor Robotic Surgery Simulator. *Journal of Endourology*, 30(3), pp.338–346.

Wiener, S. et al., 2015. Construction of a Urologic Robotic Surgery Training Curriculum: How Many Simulator Sessions Are Required for Residents to Achieve Proficiency? *Journal of endourology / Endourological Society*, 29(11), pp.1289–1293.

Wiens, P.D. et al., 2013. Using a standardized video-based assessment in a university teacher education program to examine preservice teachers knowledge related to effective teaching. *Teaching and Teacher Education*, 33, pp.24–33.

Wilkinson, J.R. et al., 2008. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical education*, 42(4), pp.364–373.

Williams, R.G. et al., 2012. A template for reliable assessment of resident operative performance: assessment intervals, numbers of cases and raters. *Surgery*, 152(4), pp.517–24– discussion 524–7.

Williams, R.G. et al., 2017. How Many Observations are Needed to Assess a Surgical Trainee's State of Operative Competency? *Annals of Surgery*, Publish Ahead of Print, p.1.

Williams, R.G. et al., 2015. Is a Single-Item Operative Performance Rating Sufficient? *Journal of surgical education*, 72(6), pp.e212–7.

Williams, R.G. et al., 2014. The Measured Effect of Delay in Completing Operative Performance Ratings on Clarity and Detail of Ratings Assigned. *JSURG*, 71(6), pp.e132–e138.

Williams, R.G., Kim, M.J. & Dunnington, G.L., 2016. Practice Guidelines for Operative Performance Assessments. *Annals of Surgery*, pp.1–15.

Wilson, A.B., Torbeck, L.J. & Dunnington, G.L., 2015. Ranking Surgical Residency Programs: Reputation Survey or Outcomes Measures? *Journal of surgical education*, 72(6), pp.e243–50.

Whitehead, C., 2012. Will the Triple C curriculum produce better family physicians? No. *Canadian family physician Médecin de famille canadien*, 58(10), pp.1071–1073–1075–8.

Xu, S. et al., 2016. Face, content, construct, and concurrent validity of a novel robotic surgery patient-side simulator: the Xperience™ Team Trainer. *Surgical Endoscopy And Other Interventional Techniques*, 30(8), pp.3334–3344.

Yamany, T. et al., 2015. Effect of postcall fatigue on surgical skills measured by a robotic simulator. *Journal of endourology / Endourological Society*, 29(4), pp.479–484.

Yanes, A.F. et al., 2016. Observation for assessment of clinician performance: a narrative review. *BMJ Quality & Safety*, 25(1), pp.46–55.

Yang, K. et al., 2016. Effectiveness of an Integrated Video Recording and Replaying System in Robotic Surgical Training. *Annals of Surgery*, pp.1–6.

Yang, K., Perez, M., et al., 2017. “Alarm-corrected” ergonomic armrest use could improve learning curves of novices on robotic simulator. *Surgical Endoscopy And Other Interventional Techniques*, 31(1), pp.100–106.

Yang, K., Zhen, H., et al., 2017. From dV-Trainer to Real Robotic Console: The Limitations of Robotic Skill Training. *Journal of surgical education*.

Young, K.A. et al., 2017. Characterizing the Relationship Between Surgical Resident and Faculty Perceptions of Autonomy in the Operating Room. *Journal of surgical education*, 74(6), pp.e31–e38.

Yudkowsky, R. et al., 2014. A Patient Safety Approach to Setting Pass/Fail Standards for Basic Procedural Skills Checklists. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 9(5), pp.277–282.

Yule, S. et al., 2015. Coaching Non-technical Skills Improves Surgical Residents' Performance in a Simulated Operating Room. *Journal of surgical education*, 72(6), pp.1124–1130.

Zevin, B., Aggarwal, R. & Grantcharov, T.P., 2012. Volume-Outcome Association in Bariatric Surgery. *Annals of Surgery*, 256(1), pp.60–71.

Zorn, K.C. et al., 2009. Training, credentialing, proctoring and medicolegal risks of robotic urological surgery: recommendations of the society of urologic robotic surgeons. *The Journal of Urology*, 182(3), pp.1126–1132.

20 Appendices

Appendix-1 Summary of included studies assessing technical skills in robotic surgery

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Aghazadeh MA et al. (2016)	17 R, 1 F, 3 S	Simulator, OR	dVSS, expert surgeons	dVSS metrics, Fundamental Inanimate Robotic Skills Tasks (GEARS)	11.5
Alemozaffar M et al. (2014)	10 novice, 10 expert	Wet (porcine)	Expert examiner	Time, OSATS	12.5
Alzahrani et al. (2013)	30 novice, 12 intermediate, 6 expert	Simulator	dVSS	dVSS metrics	12.5
Amirian et al. (2014)	26 MS	Simulator, Dry	Simbionix Suturing Module (SSM) on the dVSS	Time, Accuracy, End-product	13
Arain NA et al. (2012)	47 R, 3 F, 5 S	Dry	Expert examiner	Modified Fundamentals of Laparoscopic Surgery	13
Balasundaram I et al. (2008)	2 S, 10 R	Simulator	SEP Robotic SImulator	Error scores (Needle manipulation, suturing without traction, suturing with	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
				traction, abstract square knot, interrupted suturing)	
Brown K et al. (2017)	26 R	Simulator, Dry	dVSS	MScore	12
Chandra V et al. (2010)	20 novice, 9 expert	Simulator	ProMIS	Total task time, instrument path length, and smoothness	13.5
Chowriappa et al. (2013)	15 novice (MS, R, F), 12 expert (S)	Simulator	Computer-based	Robotic Skills Assessment Score (RSA-Score)	14.5
Chowriappa et al. (2015)	22 R, 30 F	Simulator, Dry	Expert surgeon	GEARS, urethrovesical anastomosis evaluation score	13.5
Davis JW et al. (2010)	3 R, 4 F	OR (prostatectomy)	Expert surgeon	Time to complete procedure step, quality of results relative to staff, end-product score	14
Dubin et al. (2017)	42 R, 13 F, 10 S	Simulator	dV-Trainer, dVSS, C-SATS	dV-Trainer metrics, dVSS metrics, GEARS	11.5
Dulan G et al. (2012)	8 S, 4 MS	Dry Lab	Unknown	Time, number of errors	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Finnegan et al. (2012)	18 novice, 8 intermediate, 13 expert	Simulator	dVSS	dVSS metrics	13.5
Foell et al. (2013)	29 R, 16 F, 8 S	Simulator	dVSS	dVSS metrics, time/number of errors	14
Foell K et al. (2013)	19 R, 15 F, 3 S	Dry	Expert surgeon	Time to completion, number of errors (dropped objects, collisions, excessive force, missed targets)	12.5
Gavazzi et al. (2011)	18 novice, 12 expert	Simulator	SEP Robotic Simulator	Time to completion, instrument tip trajectory, error	12.5
Ghani KR et al. (2016)	12 S	OR (prostatectomy)	Expert surgeon, crowd	GEARS, RACE	12.5
Goh AC et al. (2012)	25 R, 4 S	OR (prostatectomy)	Expert surgeon, Expert examiner, Operator	GEARS	12.5
Goh AC et al. (2015)	71 R, 4 F, 21 S	Dry	Expert examiner	Fundamentals Inanimate Robotic Skills Tasks	12
Goldenberg et al. (2017)	1 surgeon, 24 patients	OR (prostatectomy)	Expert reviewer	GEARS, generic error rating tool, continence status	11.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Gomez ED et al. (2016)	8 novice, 5 expert	Dry	Expert surgeon	OSATS GRS, GEARS	12.5
Hassan et al. (2015)	32 MS, 8 R	Simulator, Dry	dVSS, expert reviewer	Time to completion, drops, instrument collisions, instruments out of view, excessive force	13.5
Hinata N et al. (2013)	15 novice, 6 expert	Box-trainer	Expert examiner	Time to completion, technical errors	12
Holst D et al. (2015)	3 R, 2 S	Dry	Expert surgeon, crowd	GEARS	11.5
Holst D et al. (2015)	12 T (different skill levels)	Wet (porcine)	Expert surgeon, C-SATS	GEARS	12.5
Hung AJ et al. (2011)	63 (16 novice, 32 intermediate, 15 experts)	Simulator	dVSS	dVSS Metrics	12.5
Hung et al. (2012)	24 novices	Simulator	Expert Surgeons, dVSS	GOALS, dVSS Metrics	13.5
Hung AJ et al. (2013)	38 R, 11 S	Box-trainer	Expert surgeon, dVSS	GEARS, dVSS metrics	13.5
Hung AJ et al. (2015)	15 novice, 13	Simulator	dV-Trainer, expert examiner	GEARS, Time, Errors	14.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
	intermediate, 14 expert				
Hung AJ et al. (2017)	11 R, 10 F	OR (prostatectomy, partial nephrectomy)	Expert surgeon	Proficiency score, GEARS	12
Hussein et al. (2017)	28 T (R. F), 28 S	OR (prostatectomy)	Expert examiner	Prostatectomy Assessment and Competence Evaluation	13.5
Jonsson et al. (2011)	18 S, 6 R	Dry Lab	ProMIS Simulator	Time, Distance, Smoothness	10.5
Kang et al. (2014)	10 R, 10 S	Simulator	dV-Trainer	dV-Trainer metrics	12.5
Kelly et al. (2012)	19 novice (MS, R), 9 intermediate (R, F, S), 9 expert (R, F, S)	Simulator	Self-rating, dVSS	dVSS metrics	12.5
Kenney et al. (2009)	19 novice (MS, R, S), 7 expert (S)	Simulator	dV-Trainer	dV-Trainer metrics	12.5
Kim JY et al. (2015)	8 R, 3 F	Dry	Expert surgeon	Time, end-product rating score	13.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Korets R et al. (2011)	16 R	Simulator	dVSS	OSATS	13.5
Kumar R et al. (2012)	6 novice, 2 expert	Simulator	Support Vector Machines (Artificial Intelligence)	Instrument motion, telemetry, and video	14
Lee GI et al. (2017)	32 T (R, F)	Simulator, Dry	dVSS	dVSS metrics	12.5
Lee JY et al. (2012)	8 R, 2 F, 10 S	Simulator, Dry	dV-Trainer	Time, number of errors	11.5
Lendvay TS et al. (2008)	9 R, 6 S	Simulator	dV-Trainer	Time to completion, economy of motion, peak ring strain, instrument collisions, instruments out of view, time master telemanipulators were out of center	12.5
Lendvay TS et al. (2013)	27 R, 24 S	Simulator, Dry	dVSS	Time, tool path length, economy of motion, technical and cognitive errors	15
Lerner MA et. al (2010)	12 MS, 11R, 1F	Simulator, Dry	dV-Trainer	dV-Trainer metrics	12.5
Liss et al. (2015)	20 novice, 18 expert	Simulator	dVSS	MScore	13

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Liss MA et al. (2012)	6 MS, 7 R, 6 F, 13 S	Simulator	dV-Trainer, dVSS	Overall score, economy of motion, time to completion, excessive instrument force, instrument collisions, instruments out of view, master workspace range, drops	12.5
Liu M et al. (2017)	3 novice, 6 intermediate, 6 expert	Wet (swine)	Expert surgeon, expert rater	ARCS (dexterity with multiple wristed instruments, optimizing field of view, instrument visualization, optimizing master manipulator workspace, force sensitivity and control, basic energy pedal skills)	13.5
Lovegrove C et al. (2016)	5 R, 3 F, 7 S	OR (prostatectomy)	Expert surgeon	RARP assessment score	13.5
Lyons C et al. (2013)	25 novice, 8 intermediate, 13 expert	Simulator	dVSS	Overall score, economy of motion, time to complete exercise, instrument collisions, master workspace range, critical errors, instruments out of view, excessive force, missed targets, drops, misapplied energy time	12.5
Mark JR et al. (2014)	7 R	Simulator	dVSS	Epworth Sleepiness Scale, time to completion, economy of motion, instrument collisions, excessive instrument force, instruments out of view, master workspace range, misapplied energy time, needle drops, missed targets	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
McDonough et al. (2011)	10 novice, 10 expert	Simulator	ProMIS	Time to completion, path length, economy of motion	12.5
McVey R et al. (2016)	11 R, 21 F	Box-Trainer	Expert surgeon	Time to completion, number of errors, GRS	14
Meier M et al. (2016)	25 novice, 3 expert	Simulator	dVSS	dVSS metrics	13.5
Menhadji A et al. (2013)	39 R	Dry	Expert examiner	Amount of task accomplished, how accurately skill was performed	11.5
Mills et al. (2017)	10 S	Simulator	dVSS, expert surgeon	dVSS metrics, GEARS	12.5
Noureldin YA et al. (2016)	9 R	Simulator	dVSS	dVSS metrics	13
Perrenot et al. (2012)	8 novice, 6 intermediate, 5 expert, 37 no robotic experience, 19 non-physician	Simulator, Dry	dV-Trainer, expert surgeon	dV-Trainer metrics	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Phé V et al. (2017)	14 novice, 14 expert, 11 non-physician	Dry	Expert surgeon	Modified OSATS (gentleness, tissue exposure, instrument handling, time and motion, flow of operation)	12
Powers MK et al. (2016)	14 T (R, S)	OR (partial nephrectomy)	Expert surgeon, crowd-sourcing	GEARS	13.5
Raison N et al. (2017)	102 R, 121 S	Simulator	dV-Trainer	dV-Trainer metrics	14.5
Ramos P et al. (2014)	24 novice, 12 expert	Simulator, Dry	Expert surgeon	GEARS	12.5
Rashid HH et al. (2006)	2 R	OR (prostatectomy)	Expert surgeon	OSATS	11.5
Raza SJ et al. (2015)	10 novice, 10 advanced, 8 expert	Dry	Expert surgeon	Robotic Anastomosis Competency Evaluation (RACE), GEARS	13.5
Ruparel RK et al. (2014)	27 R	Simulator	dV-Trainer	Final score, economy of motion, time to completion	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Schommer E et al. (2017)	34 R	Simulator	dV-Trainer	Overall score, economy of motion, time to completion	14.5
Seixas-Mikelus et al. (2010)	6 novice, 24 expert	Simulator	Self-rate	'Fidelity' questionnaire only	7
Sethi et al. (2009)	15 novice (MS, R), 5 expert (F, S)	Simulator	dV-Trainer	dV-Trainer metrics	10
Shamim Khan M et al. (2013)	33 R	Simulator	SEP Robotic Simulator	SEP Metrics	13.5
Siddiqui NY et al. (2014)	83 R, 9 F, 13 S	Dry	Expert examiner	R-OSATS	13.5
Song PH et al. (2016)	10 MS	Dry	dVSS	dVSS metrics	11.5
Song X et al. (2016)	27 laparoscopic and robotic surgeons	Simulator	Simulator, 2 evaluators (expertise/training not defined)	Xperience Team Trainer (XTT), GOALS	14.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Stegemann et al. (2013)	8 S, 10 F, 26 R, 9 MS	Dry Lab	Trained expert raters	Errors and number of camera/clutch movememnts	14.5
Tarr ME et al. (2014)	99 R	Dry	Expert examiner	OSATS	15
Tausch et al. (2012)	5 R, 5 S	Dry	Computer-based	Time to completion, path length, economy of motion, drops, suturing and tying intracorporeal knot	11.5
Tunitsky E et al. (2013)	8 R, 4 F, 9 S	Dry	Expert surgeon	GOALS	12.5
Vernez SL et al. (2017)	25 MS	Simulator	Expert surgeon, crowd	OSATS, GEARS, GOALS	13.5
Vlaovic PD et al. (2008)	101 T (F, S)	Dry	Expert examiner	OSATS	14
Volpe A et al. (2015)	3 R, 5 F, 2 S	Simulator, OR (prostatectomy)	dVSS, expert surgeon	dVSS metrics, GEARS	12.5
White LW et al. (2015)	25 R, 24 S	Dry	Expert surgeon, C-SATS	GEARS	12.5

Study	Participants	Setting of Assessment	Raters	Measurement Tool	MERSQI
Whitehurst et al. (2015)	7 R, 8 F, 5 S	Simulator, Dry, Wet (porcine)	Expert surgeon	GEARS	14
Whittakers G et al. (2016)	20 MS, 13 R, 13 S	Simulator	RobotiX Mentor	RobotiX Mentor metrics	12.5
Wiener S et al. (2015)	16 R	Simulator	dVSS	dVSS metrics	12.5
Yamany T et al. (2015)	13 R	Simulator	dVSS	Time	14
Yang K et al. (2017)	20 MS	Simulator	dV-Trainer	dV-Trainer metrics, armrest load	13.5
Yang K et al. (2017)	40 novice, 5 experts	Simulator	dV-Trainer, dVSS	Mscore, workspace range, armrest load	12

MS – medical student, R – resident, F – fellow, S- staff, T – trainee

Appendix-2 Global Evaluative Assessment of Technical Skills (GEARS)

Depth perception				
1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual dexterity				
1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure
Efficiency				
1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression
Force sensitivity				
1	2	3	4	5
Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Autonomy				
1	2	3	4	5
Unable to complete entire task, even with verbal guidance		Able to complete task safely with moderate guidance		Able to complete task independently without prompting
Robotic control				
1	2	3	4	5
Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant

Appendix-3 Generic Error Rating Tool (GERT)

Video code			Rater code		
Surgical task group	Error mode	Time of observation	Total number	Event (description/time)	Mechanism of event
Abdominal access	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Use of retractors	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Use of energy devices	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Grasping and dissection	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Cutting, transection and stapling	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Clipping	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Suturing	Too much force/distance				
	Too little force/distance				
	Wrong orientation				
	Inadequate visualization				
Use of suction	Too much force/distance				
	Inadequate visualization				
Other unclassified	Description/time:				

Appendix-4 Standard Setting User Interface

1. Input Values

Patient characteristics	
AGE	65
BMI	27
Nerve Spare	BILATERAL
VOLUME	60
STAGE	T2c
GLEASON	7
PSA	10

Probability of Continence at 3 Months	50.0%

Highest probability possible is 82.0%

Probability of Positive Surgical Margin	20.0%

Lowest probability possible is 5.5%

Expected score for each step, by score	GEARS	PACE	PACE
Bladder Neck	4.60	4.00	
Neurovascular Bundle & Prostatic Pedicle	4.40		
Apical Dissection	4.40		4.00
Urethrovesical Anastomosis	4.00		
Needle Entry (UVA)		4.00	
Needle Driving (UVA)		4.00	
Seminal Vesicle Dissection			4.00
Posterior Dissection			4.00

2. Output Values

Minimum composite score needed

for continence outcome		for PSM outcome
GEARS score	PACE score	PACE score
4.14	4.09	4.05

Minimum score by individual step	GEARS	PACE	PACE
Bladder Neck	3.97	4.67	
Neurovascular Bundle & Prostatic Pedicle	3.57		
Apical Dissection	3.74		4.14
Urethrovesical Anastomosis	3.63		
Needle Entry (UVA)		4.21	
Needle Driving (UVA)		4.25	
Seminal Vesicle Dissection			4.19
Posterior Dissection			4.11

21 Copyright Acknowledgements

Chapter 10.5.5.1

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Apr 13, 2019

This Agreement between Dr. Mitchell Goldenberg ("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	4567330810220
License date	Apr 13, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	BJU International
Licensed Content Title	Implementing assessments of robot-assisted technical skill in urological education: a systematic review and synthesis of the validity evidence
Licensed Content Author	Mitchell G. Goldenberg, Jason Y. Lee, Jethro C.C. Kwong, et al
Licensed Content Date	Apr 24, 2018
Licensed Content Volume	122
Licensed Content Issue	3
Licensed Content Pages	19
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	Technical Skills Assessment in Robotic Surgery: Using Patient Outcomes to Set the Standard

Chapter 10.6.1

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Apr 13, 2019

This Agreement between Dr. Mitchell Goldenberg ("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	4567330908084
License date	Apr 13, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	British Journal of Surgery
Licensed Content Title	Systematic review to establish absolute standards for technical performance in surgery
Licensed Content Author	M. G. Goldenberg, A. Garbens, P. Szasz, et al
Licensed Content Date	Sep 30, 2016
Licensed Content Volume	104
Licensed Content Issue	1
Licensed Content Pages	9
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	Technical Skills Assessment in Robotic Surgery: Using Patient Outcomes to Set the Standard

Chapter 12

Chapter 13