High Stakes Technical Skill Assessments in Surgery: Development, Implementation and Predicting Performance

by

Sandra de Montbrun

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

Institute of Medical Science
University of Toronto

© Copyright by Sandra de Montbrun 2017
High Stakes Technical Skills Assessments in Surgery:
Development, Implementation and Predicting Performance

Sandra Lynn de Montbrun, MEd, M.D.
Doctor of Philosophy in Medical Science
Institute of Medical Science
University of Toronto
2017

Abstract

A paradigm shift in surgical training from an apprenticeship to a competency based model requires valid assessment platforms to assess surgical competence. Several domains of competence are currently formally assessed, however technical skill, despite being the hallmark of a surgeon, is not directly or formally assessed upon completion of surgical training. The development, evaluation and implementation of objective tools to evaluate technical skill are essential, as surgical bodies mandate the documentation of competence.

The first manuscript describes the first key step in developing a high stakes examination, using a Delphi methodology to develop a blueprint for the General Surgery Objective Structured Assessment of Technical Skill (GOSATS) exam.

An essential component of a high stakes exam is setting a pass score. The second and third manuscripts address this aspect of high stakes exam development. The second manuscript applies and compares three standard setting methodologies to the world’s largest Objective Structured Assessment of Technical Skill (OSATS) database to set passing scores for the
OSATS exam. General surgery residents (n=133) from the University of Toronto between 2002 and 2012 were then assigned a pass/fail status allowing for comparison of the consistency of the pass/fail decision across the methodologies.

The third manuscript investigates the validity of these set pass scores by evaluating the predictive ability of the OSATS pass/fail status. The general surgery residents that had passed the OSATS were compared to the residents that had failed the OSATS in terms of their PGY2 and PGY4 technical skill. The results indicate that passing the OSATS exam predicts technical skill in the PGY2 and PGY4 year.

The final manuscript is a validation study investigating the results of the first North American technical skills certification examination. In 2014 the Colorectal Objective Structured Assessment of Technical Skill (COSATS) was a mandatory component of the American Board of Colon and Rectal Surgeons certification process and represents the last paper in this dissertation. The results indicate that individuals who failed the COSATS, passed both the oral and written ABCRS exams, suggesting that the current certification process may fail to identify individuals with technical deficiencies.
Acknowledgments

I would like to thank my research supervisor Dr. Teodor Grantcharov, for his ongoing support, enthusiasm and advice throughout the past few years. Teodor, you have allowed me to grow and develop as a researcher; you will always be a mentor as I continue the pursuit of research and academia. I would also like to thank Drs. Najma Ahmed and Charlotte Ringsted, my program advisory committee, for their time, expertise and mentorship throughout my PhD.

This PhD would not have been possible without the unconditional support and encouragement of my husband Landy de Montbrun, and parents, Peter and Trudy Tschoepe.

This PhD is dedicated to my two beautiful and brilliant daughters, Devon and Reese de Montbrun. I hope that one day mom’s accomplishments will make you proud, and inspire you to reach for the stars; you can accomplish anything you put your mind to. Your mom loves you more than words can say.
Contributions

Sandra de Montbrun independently prepared this thesis and all aspects of the included original research studies from: study design, data collection, analysis and writing. This thesis contains four original manuscripts to which Sandra de Montbrun is the primary author. All contributions by coauthors are described in detail below:

Dr. Peter Szasz - design, analysis and manuscript preparation for Chapters 3.

Dr. Marisa Louridas - design, analysis and manuscript preparation for Chapters 3 and 5.

Dr. Helen MacRae – design, manuscript preparation for Chapter 6.

Anton Svendrovski – statistical assistance for Chapters 4, 5 and 6.

Lisa Satterthwaite – data collection for Chapter 4.
# Table of Contents

Acknowledgments ........................................................................................................ iv

Contributions ............................................................................................................... v

Table of Contents ....................................................................................................... vi

List of Tables .............................................................................................................. xiii

List of Figures ............................................................................................................ xv

List of Abbreviations ................................................................................................. xvii

1 Chapter 1: Thesis Overview, Rationale, Hypothesis and Specific Aims ................... 1
   1.1 Thesis Overview ............................................................................................... 1
   1.2 Thesis Rationale ............................................................................................... 3
   1.3 Hypotheses ...................................................................................................... 4
   1.4 Research Aims ............................................................................................... 4

2 Chapter 2: Literature Review .................................................................................. 6
   2.1 Residency Training Past and Present ............................................................. 7
      2.1.1 Traditional Mentorship Model ................................................................ 7
      2.1.2 Competency Based Medical Education ...................................................... 7
         2.1.2.1 Competency Based Surgical Curriculum: The University of Toronto Orthopaedic Experience ........................................................................... 8
         2.1.2.2 Competency Based Assessment ............................................................ 9
   2.2 Technical Skill Acquisition .............................................................................. 10
      2.2.1 Fitts and Posner’s Three Stage Theory of Motor Acquisition ................ 10
      2.2.2 Dreyfus and Dreyfus’ Model of Skill Acquisition .................................... 11
   2.3 Technical Skill Assessment ............................................................................ 14
      2.3.1 Definition of Assessment ......................................................................... 14
      2.3.2 Formative versus Summative Assessment ............................................... 16
2.3.3 Theory of Technical Skill Assessment ...............................................................16

2.4 Current Methods of Assessing Technical Skill....................................................19

2.4.1 Traditional Platforms to Assess Technical Skill.................................................19

2.4.1.1 Direct Observation..........................................................................................19

2.4.1.2 Surgical Logs................................................................................................19

2.4.1.3 In Training Evaluation Reports .....................................................................20

2.4.1.4 Prescribed Time..............................................................................................20

2.4.1.5 Surrogates of Technical Skill .......................................................................21

2.4.1.6 Issues with Traditional Assessment Methods...............................................21

2.4.2 Objective Assessments of Technical Skill.........................................................21

2.4.2.1 Procedural Based Assessments (PBA).........................................................21

2.4.2.2 Objective Structured Assessment of Technical Skill (OSATS) Examination .................................................................22

2.4.3 Current Performance/Procedure Based Assessments Used in High Stakes Decisions.........................................................................................................................26

2.4.3.1 The Israeli Board of Anesthesia Examination ..............................................26

2.4.3.2 The Medical Council of Canada (MCC) Objective Structured Clinical Examination (OSCE) .................................................................28

2.4.3.3 The Fundamentals of Laparoscopic Surgery (FLS) Examination ...............29

2.4.3.4 The European Board of Vascular Surgery Examination ............................30

2.4.3.5 The Colorectal Objective Structured Assessment of Technical Skill (COSATS) Examination .................................................................30

2.5 The Current Board Certification Process ............................................................33

2.5.1 Canadian Surgical Board Certification.............................................................34

2.5.2 American Surgical Board Certification ............................................................35

2.6 Competence ........................................................................................................36

2.6.1 Definition .........................................................................................................36
3.3 Methods..................................................................................................................63

3.3.1 Delphi Methodology ..........................................................................................64

3.3.1.1 Expert Panel..................................................................................................64

3.3.1.2 Anonymity.....................................................................................................64

3.3.1.3 Delphi Rounds ..............................................................................................65

3.3.1.4 Aggregation of Results and Definition of Consensus ..................................65

3.3.2 Inclusion Criteria for GOSATS Blueprint ..........................................................66

3.3.3 Administration ...................................................................................................66

3.4 Results.....................................................................................................................66

3.4.1 Participants.........................................................................................................66

3.4.2 Consensus ..........................................................................................................67

3.4.3 Final GOSATS Exam Blueprint .........................................................................67

3.5 Discussion ...............................................................................................................72

3.6 Conclusions.............................................................................................................75

4 Chapter 4: Setting passing scores for technical performance in surgery: lessons learned from 10 years of resident assessment.................................................................76

4.1 Abstract..................................................................................................................76

4.2 Introduction.............................................................................................................78

4.3 Methods..................................................................................................................79

4.3.1 Establishing Passing Scores for Each OSATS Task..........................................80

4.3.1.1 The Contrasting Groups Methodology.........................................................82

4.3.1.2 Borderline Group Methodology .................................................................83

4.3.1.3 Borderline Regression Methodology.........................................................83

4.3.2 Establishing a Pass Criteria for the Overall Exam..............................................85

4.4 Results.....................................................................................................................85

4.4.1 Establishing Passing Scores for Each OSATS Task..........................................85
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.2</td>
<td>Establishing a Pass Criteria for the Overall Exam</td>
<td>87</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>89</td>
</tr>
<tr>
<td>4.6</td>
<td>Conclusion</td>
<td>94</td>
</tr>
<tr>
<td>5</td>
<td>Chapter 5: Passing a technical skill exam in the first year of residency predicts future performance</td>
<td>95</td>
</tr>
<tr>
<td>5.1</td>
<td>Abstract</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Introduction</td>
<td>97</td>
</tr>
<tr>
<td>5.3</td>
<td>Methods</td>
<td>98</td>
</tr>
<tr>
<td>5.4</td>
<td>Results</td>
<td>100</td>
</tr>
<tr>
<td>5.5</td>
<td>Discussion</td>
<td>100</td>
</tr>
<tr>
<td>5.6</td>
<td>Conclusion</td>
<td>102</td>
</tr>
<tr>
<td>6</td>
<td>Chapter 6: Implementing and evaluating a national certification technical skills exam: the Colorectal Objective Structured Assessment of Technical Skill (COSATS)</td>
<td>110</td>
</tr>
<tr>
<td>6.1</td>
<td>Abstract</td>
<td>110</td>
</tr>
<tr>
<td>6.2</td>
<td>Introduction</td>
<td>112</td>
</tr>
<tr>
<td>6.3</td>
<td>Methods</td>
<td>113</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Structure of the COSATS Exam</td>
<td>113</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Assessment Instruments</td>
<td>113</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Participants</td>
<td>114</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Examiners</td>
<td>114</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Exam Administration</td>
<td>117</td>
</tr>
<tr>
<td>6.3.6</td>
<td>Validity Evidence</td>
<td>117</td>
</tr>
<tr>
<td>6.3.6.1</td>
<td>Internal Structure Evidence</td>
<td>117</td>
</tr>
<tr>
<td>6.3.6.2</td>
<td>Relationship to Other Variables Evidence</td>
<td>117</td>
</tr>
<tr>
<td>6.3.6.3</td>
<td>Consequences Evidence</td>
<td>118</td>
</tr>
<tr>
<td>6.3.6.3.1</td>
<td>Borderline Group Methodology</td>
<td>118</td>
</tr>
</tbody>
</table>
### 6.3.6.3.2 Borderline Regression Methodology ........................................... 118

### 6.3.6.3.3 Pass Criteria for the COSATS .................................................. 118

### 6.4 Results ........................................................................................................... 119

#### 6.4.1 Internal Structure Evidence ................................................................. 119

#### 6.4.2 Relationship to other Variables Evidence .............................................. 119

#### 6.4.3 Consequences Evidence ......................................................................... 123

### 6.5 Discussion .................................................................................................... 125

### 6.6 Conclusion ................................................................................................... 128

### 7 Chapter 7: General Discussion and Limitations ........................................ 130

#### 7.1 Thesis Summary .......................................................................................... 130

#### 7.2 The Use of a Unified Contemporary Model of Validity in Surgical Education .......... 132

#### 7.3 Simulated High Stakes Assessment Contextualized in Miller’s Theory of Assessment .134

#### 7.4 The Rational for Using Miller’s “Shows How” Level of Assessment for High Stakes Assessment of Technical Skill ................................................................. 135

#### 7.5 Fluidity of Competence .............................................................................. 136

#### 7.6 Implementation of High Stakes Assessments for Training and Board Certification ...... 137

##### 7.6.1 Feasibility .................................................................................................. 137

###### 7.6.1.1 Cost of Taking the COSATS Examination ................................................. 138

###### 7.6.1.2 Large Scale Exam Administration ........................................................... 138

##### 7.6.2 Dealing With the Failing Residents ......................................................... 140

###### 7.6.2.1 The Appeal Process ............................................................................... 141

##### 7.6.3 Exam Security and Training to the Test .................................................... 142

#### 7.7 Limitations .................................................................................................. 143

##### 7.7.1 The Use of Simulation in High Stakes Assessment ..................................... 144

##### 7.7.2 Fluidity of Competence ........................................................................... 145
7.7.3 Simulated Model Limitations ......................................................... 145

8 Chapter 8: Future Directions ................................................................. 147

8.1 Knowledge Translation Plan ............................................................... 147

8.2 Assessing Technical Skill beyond Training and Certification ................ 149

8.3 Assessing Broader Domains of Surgical Competence ....................... 149

8.4 Developing and Implementing a Formative Assessment Program .......... 150

References .................................................................................................. 152
List of Tables

Table 1. Fitts and Posner’s three stage theory of motor skill acquisition ........................................ 11
Table 2. Dreyfus and Dreyfus five level model of skill acquisition ....................................................... 13
Table 3. Messick’s five major sources of validity evidence ................................................................. 44
Table 4. Definition of the four inferences in Kane’s validity argument .............................................. 46
Table 5. Participating general surgery programs and subspecialty of program director .................... 68
Table 6. Positive consensus on procedures that could be included in a general surgery certification examination .......................................................................................................................... 69
Table 7. Passing scores for each technical skills task for all three standard setting methods ...... 86
Table 8. Consistency of the pass/ fail status across three methodologies........................................... 89
Table 9. Comparison of PGY2 overall technical skill score between residents that passed and failed the OSATS ................................................................................................................ 101
Table 10. Comparison of PGY4 overall technical skill score between residents that passed and failed the OSATS .............................................................................................................. 102
Table 11. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a compensatory borderline group method ........................................... 121
Table 12. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a conjunctive borderline group method ............................................... 121
Table 13. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a compensatory borderline regression method ......................................... 122
Table 14. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a conjunctive borderline regression method.............................................. 122

Table 15. COSATS technical skill stations with mean scores, passing scores and passing rates
..................................................................................................................................................... 124
List of Figures

Figure 1. Miller’s framework for clinical assessment (Adapted from Miller, 1990) .................. 17

Figure 2. Objective Structured Assessment of Technical Skill (OSATS) global rating scale (Adapted from Martin et al., 1997) ........................................................................................................... 25

Figure 3. Simulated COSATS models a. ileal pouch anal anastomosis b. presacral bleeding c. coloanal anastomosis d. perianal fistula e. loop ileostomy f. virtual reality colonoscopy .......... 32

Figure 4. The Royal College of Physicians and Surgeons of Canada CanMEDS framework ..... 38

Figure 5. Schematic representation of the borderline group standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015) ........................................... 53

Figure 6. Schematic representation of the borderline regression standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015) .......... 54

Figure 7. Example of an overall performance scale (Adapted from de Montbrun S., Satterthwaite L., Roberts P., MacRae H. 2016) .......................................................................................................................... 55

Figure 8. Schematic representation of the contrasting groups standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015) .......... 57

Figure 9. OSATS Global rating scale .......................................................................................... 81

Figure 10. The contrasting groups standards setting methodology ........................................ 82

Figure 11. The borderline group standard setting methodology ............................................. 84

Figure 12. Borderline regression standard setting methodology ............................................. 84

Figure 13. Passing scores for all three standards setting methodologies .................................. 88

Figure 14. Boxplots comparing overall technical skill score at the PGY2 and PGY4 level between residents who pass and fail the OSATS exam during PGY1 using a. contrasting groups method b. borderline group method c. borderline regression method ......................... 105
Figure 15. COSATS global rating scale ................................................................. 115

Figure 16. COSATS overall performance scale ..................................................... 116

Figure 17. Scatterplot comparing COSATS score with the ABCRS oral exam score .......... 120
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCRS</td>
<td>American Board of Colon and Rectal Surgery</td>
</tr>
<tr>
<td>ABS</td>
<td>American Board of Surgery</td>
</tr>
<tr>
<td>ACGME</td>
<td>Accreditation Council for Graduate Medical Education</td>
</tr>
<tr>
<td>ASCRS</td>
<td>American Society of Colon and Rectal Surgeons</td>
</tr>
<tr>
<td>BG</td>
<td>borderline group method</td>
</tr>
<tr>
<td>BRM</td>
<td>borderline regression method</td>
</tr>
<tr>
<td>CBE</td>
<td>Competency based education</td>
</tr>
<tr>
<td>CBME</td>
<td>Competency based medical education</td>
</tr>
<tr>
<td>CG</td>
<td>contrasting groups method</td>
</tr>
<tr>
<td>COSATS</td>
<td>Colorectal Objective Assessment of Technical Skill</td>
</tr>
<tr>
<td>EBVS</td>
<td>European Board of Vascular Surgery</td>
</tr>
<tr>
<td>FITER</td>
<td>Final in training evaluation report</td>
</tr>
<tr>
<td>FLS</td>
<td>Fundamentals of Laparoscopic Surgery</td>
</tr>
<tr>
<td>GOSATS</td>
<td>General Surgery Objective Structured Assessment of Technical Skill</td>
</tr>
<tr>
<td>GRS</td>
<td>global rating scale</td>
</tr>
<tr>
<td>ITER</td>
<td>In training evaluation report</td>
</tr>
<tr>
<td>LMCC</td>
<td>Licentiate of the Medical Council of Canada</td>
</tr>
<tr>
<td>MCC</td>
<td>Medical Council of Canada</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MCCQE</td>
<td>Medical Council of Canada Qualifying Examination</td>
</tr>
<tr>
<td>MISTELS</td>
<td>McGill Inanimate System for Training and Evaluation of Laparoscopic Skills</td>
</tr>
<tr>
<td>OSATS</td>
<td>Objective Structured Assessment of Technical Skill</td>
</tr>
<tr>
<td>OSCE</td>
<td>Objective Structured Clinical Examinations</td>
</tr>
<tr>
<td>PBA</td>
<td>Procedural Based Assessments</td>
</tr>
<tr>
<td>PGY</td>
<td>postgraduate year</td>
</tr>
</tbody>
</table>
Chapter 1: Thesis Overview, Rationale, Hypothesis and Specific Aims

1.1 Thesis Overview

Chapter 1. Thesis Overview, Rationale, Hypothesis and Specific Aims

This chapter provides an overview of the dissertation, with hypotheses and aims.

Chapter 2. Literature Review

The training and assessment of technical skill during residency, both traditional and contemporary are reviewed. Theories of technical skill acquisition and technical skill assessment are outlined. The definition of “competence” is explored and defined for the purpose of this dissertation. The surgical board certification process is reviewed for both Canada and the United States. Traditional and contemporary theories of validity are reviewed to provide a foundation for guiding the validation studies within this dissertation. Finally, standard setting methodologies are described, and their applicability to performances based assessments is discussed.

Chapter 3: Developing the blueprint for a Canadian general surgery technical skills certification exam: a validation study.

Chapter 3 addresses the first specific aim of the thesis in manuscript form. This study used a Delphi consensus methodology to develop a blueprint for a technical skills certification examination for graduating Canadian general surgery residents. This is a validation study outlining the first key step in the development of a high stakes examination.
Chapter 4: Setting Passing Scores for Technical Performance in Surgery: Lessons Learned From 10 Years of Resident Assessment

Chapter 4 addresses the second specific aim of this thesis in manuscript form. This study applies and compares three different standard setting methodologies to the world’s largest Objective Structured Assessment of Technical Skill (OSATS) database. This is the first time that passing scores have been set for the OSATS exam. Initial evidence of validity is sought for these passing scores by comparing results across the three methodologies. This is the first large scale application of standard setting methodologies to performance based assessments of technical skill.

Chapter 5: Passing a technical skill exam in the first year of residency predicts future performance

Chapter 5 addresses the last specific aim of this thesis in manuscript form. While the gold standard of assessment is the ability to predict performance this study investigates the predictive ability of the pass/fail status of the OSATS exam using the passing scores set in Chapter 4.

Chapter 6, Implementing and Evaluating a National Certification Technical Skills Exam: the Colorectal Objective Structured Assessment of Technical Skill (COSATS)

Chapter 6 addresses the fourth specific aim of this thesis in manuscript form. This study outlines the implementation of the first North American board certification simulated technical skills examination. The Colorectal Objective Assessment of Technical Skill (COSATS) was implemented into the American Board of Colon and Rectal Surgery (ABCRS) certification examination in September 2014. The results of this high stakes examination are discussed.
Chapter 7: General Discussion and Limitations

Chapter 7 provides an overall discussion on the dissertation looking at how the manuscripts taken together build validity evidence for the objective assessment of technical skill. It explores the argument behind the use of simulation in high stakes examination, the issue of competence fluidity and the limitations and barriers to performance based assessment in training and certification.

Chapter 8: Future Directions

Chapter 8 explores avenues of knowledge translation, and further work to be explored including assessing skill beyond the point of certification, assessing other domains of competence and future work to be done with formative assessment of technical skill.

1.2 Thesis Rationale

Assessing technical skill of surgical residents is an essential component of residency training both during and upon completion of training. While traditionally, formal assessment of surgical residents has included written examinations and oral examinations to assess knowledge and judgment respectively, the domain of technical skill has not typically been formally or objectively evaluated either during residency or at the time of board certification.

There is little doubt that technical skill is a fundamental domain of competence for surgical specialties; however, platforms aimed at assessing technical skill have been variably implemented into training, and most have not been developed for the purpose of high stakes assessment such as matriculation or certification.
With technical skill being the hallmark of surgical specialties and with the current movement towards competency based training, the development, evaluation and implementation of technical skills assessments should be a priority.

A technical skills exam used for high stakes decisions, such as certification, requires that the tool demonstrate a high level of validity for the interpretation of exam scores. Different sources of validity are continually sought to build evidence for the interpretation of scores; validity evidence is sought from the beginning of exam development and continues all the way to the standard setting process where a passing score is defined.

This thesis focuses on addressing the current gaps in knowledge within the domain of high stakes assessment of technical skill through developing, validating and implementing high stakes technical skill assessments into surgical training.

### 1.3 Hypotheses

We first hypothesize that a Delphi methodology can be used to develop a blueprint for a high stakes assessment and build evidence of content validity for an objective assessment of general surgery technical skill. Secondly, we hypothesize that various standard setting methodologies can be applied to performance based assessments of technical skill, producing stable pass/fail decisions. Thirdly, we hypothesize that a pass/fail status on the OSATS technical skills examination has the ability to predict future technical skill. Lastly, we hypothesize that an objective assessment of technical skill can be incorporated into surgical board certification, with evidence of validity of the interpretation of examination scores, and evidence that a technical skills exam measures a different construct than the current written and oral examinations.

### 1.4 Research Aims

The overarching aim of this thesis is to investigate the various stages of the development, implementation and consequences of high stakes objective technical skills assessments for certification grounded in contemporary theories of collecting validity evidence.
Given the overarching objective and hypothesis, this thesis is structured around the following four specific objectives:

Specific Aim #1: To establish consensus on a blueprint for a technical skills certification examination for graduating Canadian general surgery residents using a Delphi methodology.

Specific Aim #2: To set, compare and evaluate three standard setting methodologies for performance based assessment of technical skill using the world’s largest Objective Structured Assessment of Technical Skill (OSATS) database.

Specific Aim #3: To build evidence of validity for the set OSATS scores by investigating the ability of the OSATS pass/ fail status to predict future technical skill.

Specific Aim #4: To evaluate the implementation of a high stakes technical skill examination into the American Board of Colon and Rectal Surgery (ABCRS) certification.
2 Chapter 2: Literature Review

Parts of this chapter will be published in the textbook entitled: Simulation for Surgery and Surgical Subspecialties. Springer. Chapter 11; Use of Simulation in High-Stakes Summative Assessments in Surgery.

Surgical competence is of growing concern to the public, as well as to regulatory and licensing bodies. There is increasing pressure to measure operative skill, set standards and ensure the competence of surgeons and trainees. (Darzi, Smith, & Taffinder, 1999) Also, with high profile surgical cases in the media, the public needs reassurance that surgeons are meeting high quality competency standards before they are allowed to practice independently. A study looking at adverse events in Canadian hospitals documented that 51.4% of adverse events were due to surgical care, and the most common type of adverse event was due to surgical procedures. (Baker et al., 2004) The interpretation of this data suggested that improving surgical safety would translate into improved patient outcomes. In the 1990s the Accreditation Council for Graduate Medical Education (ACGME) published a report outlining the required domains of surgical competence. (Mery, Greenberg, Patel, & Jaik, 2008)

Over the past several decades there has been a shift in surgical training and evaluation from an apprenticeship style Halstedian model to a competency based training framework. In parallel with this paradigm shift, simulation has become a major platform for both the training and assessment of technical skill.
2.1 Residency Training Past and Present

2.1.1 Traditional Mentorship Model

The traditional Halstedian approach to surgical training relied on the acquisition of skill and knowledge through an apprenticeship model. This model required that surgical trainees spend long hours in a hospital setting gaining adequate exposure to many cases in order to build their clinical skills. (Halpenny, 1918)

The current surgical training milieu required that this approach to surgical training be re-evaluated. Limitations on resident duty hours both within Canada (National Steering Committee on Resident Duty Hours, 2013) and the United States (Accreditation Council for Graduate Medical Education, 2011), as well as issues of resident fatigue, wellbeing and patient safety (Bolster & Rourke, 2015), have initiated a paradigm shift towards competency based training systems.

2.1.2 Competency Based Medical Education

Competency based medical education (CBME) is a paradigm shift in how we train and assess the new generation of medical trainees. (Hawkins et al., 2015) A recent systematic review has attempted to define competency based education. (Frank et al., 2010) Competency based education (CBE) has been defined as:

“...an approach to preparing physicians for practice that is fundamentally oriented to graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It deemphasizes time-based training and promises greater accountability, flexibility, and learner centeredness.”

With restricted duty hours limiting clinical exposure during surgical training, residents may not be getting sufficient exposure to reach competence in the procedures required of their specialty.
One solution to counteract the effects of limited operative exposure is to implement a competency based training model.

Competency based training models are being mandated by surgical bodies (The Royal College of Physician and Surgeons of Canada, 2014), (Accreditation Council for Graduate Medical Education, 2015) and, while the intention of these training models is to ensure that individuals have achieved competence in expected milestones throughout their residency, this system of training inevitably requires the availability of validated tools to teach and assess the achievement of these competencies.

With this ambitious endeavour comes an ever increasing need for the development and evaluation of valid tools to train and assess the modern surgical resident. Surgical education is currently grappling with how to develop these tools, how to implement and study their effects, and how to ensure their feasibility in the current residency training system.

2.1.2.1 Competency Based Surgical Curriculum: The University of Toronto Orthopaedic Experience

One of the best examples of the implementation of a competency based curriculum comes from the University of Toronto’s orthopaedic training program. This competency based education program is based on the concept that residents will graduate from their program when they are competent, not after they have completed a prescribed period of time. Issues of patient safety, resident fatigue and resident duty hours were all an impetus for this residency training reform. Recently, the three-year experience of this program has been published (Ferguson et al., 2013) While the long term results of this program are still uncertain, the first three years of data suggest that residents in the competency based curriculum achieve technical skill equivalent to traditionally trained residents in an expedited fashion (Sonnadara et al., 2012) and slightly outperform traditional trainees in terms of knowledge.
The feasibility of implementing this type of training system across all surgical specialties on a national level remains to be established. While the majority of surgical residencies continue to take the traditional time based approach to surgical training, competency based assessments are being incorporated throughout surgical training.

2.1.2.2 Competency Based Assessment

Over the past several decades tools to assess technical competence have been developed and evaluated in the literature. (J. Beard, Rowley, Bussey, & Pitts, 2009; S. L. de Montbrun et al., 2013; Martin et al., 1997; Vassiliou et al., 2006) While valid and reliable assessment platforms are essential components of competency based programs, thus far many of the available tools lack an essential step in their development. In order to differentiate between a competent and non-competent performance, a standard of performance in the form of a passing score needs to be established. Without this threshold of performance being defined, a score alone on an assessment tool has no meaning and cannot be translated into a judgment of competence. The currently available tools have for the most part failed to set passing scores and thus limit their ability to make decisions of competence. Standard setting methodologies are methodologies used to set passing scores and, while they have a relatively long history in written examinations, their use in performance based assessments in general has been less well studied. Their application to technical skill assessment platforms is extremely limited. This is an essential component of competency based technical skills assessment that needs to be investigated and incorporated into competency based surgical training.

A recent systematic review investigating the currently available methodologies for assessing technical competence identified several tools that are commonly used in surgical training to assess technical skill. These included Likert scales, benchmarks, binary outcomes, novel tools and surrogate outcomes. (Szasz, Louridas, Harris, Aggarwal, & Grantcharov, 2014) While the authors acknowledge the availability of a multitude of tools to assess technical competence, they highlight the lack of use of standard setting methodologies to set passing scores for these tools, limiting their ability to clearly differentiate between a competent and non-competent
performance. Although standard setting is an essential component of competency based programs, standard setting has been grossly understudied in the domain of surgical assessment, and is an area that requires further investigation.

2.2 Technical Skill Acquisition

While surgeons need to develop many domains of competence, technical skill is the hallmark of all surgeons. Becoming a safe and independent surgeon requires the attainment of technical skill. Understanding how individuals gain technical skill can be understood through established theories of skill acquisition. Two established theories of motor acquisition include Fitts and Posner’s three stage theory of motor acquisition, and Dreyfus and Dreyfus’ five stage model of skill acquisition.

2.2.1 Fitts and Posner’s Three Stage Theory of Motor Acquisition

Fitts and Posner in 1967 describe their now classic theory of motor skill acquisition. (Fitts & Posner, 1967) Their model outlines three sequential stages in the acquisition of skill. These three stages are outlined in Table 1. The cognitive stage is the first stage and involves the individual understanding the goals of the task. During this phase, movements are largely controlled consciously and are typically slow, inefficient and inconsistent. The associative stage is the second stage and is characterized by the need for less cognitive activity as movements become more fluid, efficient and reliable. While there is still thinking taking place, the executed task is becoming more automated. The final stage is automation, when movements become largely automated. This stage demonstrates accurate, efficient and consistent movement, and requires little cognitive effort. The task of learning how to tie a surgical knot can be understood using Fitts and Posner’s theory. The medical student learning to tie a surgical knot would be in the cognitive stage, requiring a significant amount of cognition as they learn how to hold the suture, and how to move their fingers and hands to execute the task. A resident exemplifies the associative phase, as movements become fluid with less cognitive effort, but there is still room
for perfecting their tying as they continually practice throughout their training. The staff surgeon would be in the autonomous phase, where the movements become so automatic that breaking down knot tying into its individual steps can be challenging.

Table 1. Fitts and Posner’s three stage theory of motor skill acquisition

<table>
<thead>
<tr>
<th>Stage of Motor Skill Acquisition</th>
<th>Characteristics of Performance</th>
<th>Cognitive Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>Erratic, slow, distinctive steps</td>
<td>High</td>
</tr>
<tr>
<td>Associative</td>
<td>Fluid, reliable, efficient</td>
<td>Less cognitive effort</td>
</tr>
<tr>
<td>Autonomous</td>
<td>Accurate, consistent, efficient</td>
<td>Little/ no cognitive effort</td>
</tr>
</tbody>
</table>

(Adapted from Fitts & Posner, 1967)

2.2.2 Dreyfus and Dreyfus’ Model of Skill Acquisition

In 1986 Dreyfus and Dreyfus described a five stage model of adult skill acquisition that can be applied to understanding how residents acquire technical skill. (Dreyfus, 2004) Like Fitts and Posner’s model, Dreyfus and Dreyfus’ model also focuses on the development of automaticity as individuals advance through the five stages (novice, advanced beginner, competence, proficiency, expert). The five stages and their characteristics are outlined in
Table 2, and describe the progression from a rigid rule based understanding of a skill to an intuitive mode of reasoning and automaticity.

This model is helpful in conceptualizing performance at different levels of skill acquisition. Dreyfus and Dreyfus’ model differentiates between the terms “competent” and “proficient”. While the surgical literature often uses these terms interchangeably, it is important to recognize that these stages are distinct and represent different stages of learning. This model can guide the development of an assessment tool, recognizing what the tool is intended to measure, whether it be “competency” or “proficiency”.

Understanding how residents acquire technical skill using either the Fitts and Posner or Dreyfus and Dreyfus model of skill acquisition can help guide the implementation of appropriate assessment tools for the level of skill that is intended to be measured. For example, George Miller’s framework for assessing clinical competence also used a novice to expert progression to guide the use of appropriate assessment tools for various levels of clinical competence. (Miller, 1990)
Combining models of skill acquisition with models of assessment allows for the appropriate matching of skill level with assessment tool.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Characteristics</th>
<th>Autonomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>Adherence to rules or plans</td>
<td>Needs close supervision and instruction</td>
</tr>
<tr>
<td></td>
<td>Little situational perception</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No discretionary judgment</td>
<td></td>
</tr>
<tr>
<td>Advanced Beginner</td>
<td>Action based on attributes or aspects</td>
<td>Able to achieve some steps using own judgment, but supervision needed for overall task</td>
</tr>
<tr>
<td></td>
<td>Situational perception still limited</td>
<td></td>
</tr>
<tr>
<td>Competent</td>
<td>Has standardized and routinized procedures</td>
<td>Able to achieve most tasks using own judgment</td>
</tr>
<tr>
<td><strong>Proficient</strong></td>
<td>Sees what is most important in a situation</td>
<td>Able to take full responsibility for own work, and coach others</td>
</tr>
<tr>
<td></td>
<td>Perceives deviations from the normal pattern</td>
<td></td>
</tr>
<tr>
<td><strong>Expert</strong></td>
<td>No longer relies on rules or guidelines</td>
<td>Goes beyond existing standards and creates own interpretations</td>
</tr>
<tr>
<td></td>
<td>Intuitive decision making</td>
<td></td>
</tr>
</tbody>
</table>

(Adapted from Dreyfus, 2004)

2.3  Technical Skill Assessment

2.3.1  Definition of Assessment

*The Standards for Educational and Psychological Testing* defines assessment as:

>“Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects or programs.” ([American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999](#))

Different platforms of assessment are available to measure different constructs. The Accreditation Council for Graduate Medical Education and American Board of Medical Specialties have outlined the available assessment methods for resident evaluation. ([Accreditation Council for Graduate Medical Education & American Board of Medical Specialties, 2000](#)) They have described: 360-degree evaluation instruments, chart stimulated
recall oral examinations, checklist evaluation of live or recorded performances, global rating of live or recorded performances, Objective Structured Clinical Examinations (OSCE), procedure-operative/case logs, patient surveys, portfolios, record reviews, simulations and models, standardized oral examinations and standardized patient examinations. This broad list of assessment modalities demonstrates the variety of tools that are available to measure different domains of competence. When selecting an assessment tool for resident evaluation it is imperative to keep in mind the construct or competency that is being assessed as well as the purpose and goal of the assessment (e.g. selection into residency, progression through training, certification) in order to select an appropriate tool.

The purpose of assessment is to quantify or measure this underlying construct. (Andreatta & Gruppen, 2009) Constructs have been defined as:

“Intangible collections of abstract concepts and principles which are inferred from behavior and explained by educational and psychological theory” (Downing, 2003)

Constructs within surgery would include theoretical entities such as knowledge, judgment and technical skill. Although these entities can be described and do exist, they can be difficult to measure. The construct being measured with an assessment tool needs to be clearly defined, to help guide the development of the assessment tool.

While surgeons need to develop many competencies, this dissertation focuses on the construct of technical skill. It is difficult to perfectly isolate and assess a single construct, and it is recognized that assessing technical skill also inevitably requires the possession of other constructs such as knowledge. For example, in asking a resident to perform the technical task of “sigmoid colon mobilization”, the resident would need to have “knowledge” of the steps of the procedure. The same would hold true for an elementary math test. While the purpose of the test is to assess math, in order to answer the question a child would need to know how to read, and thus the test secondarily assesses reading ability.
2.3.2 Formative versus Summative Assessment

Assessment can be broadly classified as either formative or summative. Formative assessments are used for the purpose of providing feedback, advancing learning and allowing for deliberate practice. They are usually completed throughout the course of study in order to allow students to identify areas of weakness, and to promote achievement and learning in these identified area of weakness.(Downing & Haladyna, 2009)

On the other hand, summative assessments are used for making decisions in high stakes examinations, such as certification and licensure. They are used to assess the achievement of knowledge or ability upon completion of a course, as a final measure of students’ learning. While they may provide students with feedback, the main purpose or goal of summative assessments is to provide a final evaluation on the achievement of curricular goals.

This distinction is important when considering the development of an assessment tool. The goal of the assessment needs to be at the forefront of development, as the level of evidence for validation and psychometric properties of the tool will vary greatly between a summative high stakes examination and a formative assessment tool being developed for learning purposes.(Downing, 2003)

2.3.3 Theory of Technical Skill Assessment

Just as there are frameworks to help understand the acquisition of technical skill, frameworks have also been developed to understand assessment. Miller’s pyramid serves as a useful taxonomy for understanding assessment within surgical education. This pyramid can parallel theories of skill acquisition, as they are complementary in their understanding of skill progression.

Physician competence is a complex construct and no single assessment tool can establish achievement of “competence”. Various assessment tools assess different levels of skill, and Miller’s pyramid illustrates a graduated assessment of physician skill. This graduated
assessment parallels Fitts and Posner’s and Dreyfus and Dreyfus models of skill acquisition. The complexity of skill assessment is highlighted by Miller, who says:

“…no single assessment method can provide all the data required for judgment of anything so complex as the delivery of professional services by a successful physician.” (Miller, 1990)

Miller’s framework is conceptualized as a pyramid and outlined in Figure 1.

![Figure 1. Miller’s framework for clinical assessment (Adapted from Miller, 1990)](image)

In the base of the pyramid sits cognitive knowledge. This knowledge serves as the base upon which other health care knowledge and skill would be developed. At this level of the pyramid, labeled as “Knows”, students demonstrate factual knowledge, the basic principles upon which more complex learning can build. (Downing & Haladyna, 2009) This level of assessment is exemplified by written tests.
The next level of the pyramid is “Knows how”, where individuals not only possess foundational knowledge but can also use the knowledge to interpret data, understand relationships among concepts, create a differential diagnosis and develop a management plan. This level hallmarks a more cognitively complex understanding. Miller would classify this level as “competent”. Oral examinations, aimed at assessing clinical judgment, exemplify assessment at this level of Miller’s pyramid.

“Shows how” is the next level in Miller’s Pyramid. This level aims at assessing if students are able to demonstrate or “show” what they “know how” to do. This level typically involves performance based assessments. Examples at this level would include standardized patients or simulated technical skill assessments in a skills laboratory. While these assessments may be somewhat artificial, they elevate the assessment beyond just knowledge.

The peak of the pyramid represents the highest level of assessment. This level represents “Does”, and assesses skill in a real world setting with real patients. This is the most authentic uncued professional assessment. An example of assessment at the “Does” level would be the observation and expert evaluation of a resident’s performance during a surgical case.

This framework is useful for matching the appropriate assessment methods with the competency being tested. For example, within surgical training, assessing a medical student’s knowledge of grading hemorrhoids would be best accomplished through a written examination. Assessing a junior resident’s conceptualization of a differential diagnosis and management plan for hemorrhoids would be best achieved with a structured oral examination. Assessing the competence of a senior resident in performing a hemorrhoidectomy would be best done either using a simulated model in the surgical skill laboratory or, more authentically, by assessing them performing a hemorrhoidectomy on a real patient. Matching the intention of the assessment with the type of assessment is essential.
2.4 Current Methods of Assessing Technical Skill

A multitude of tools have been used to assess the technical skill of surgical residents. These different modalities vary in terms of their rigorousness in development and available literature on validation and reliability. Individual residency programs across the globe vary in terms of the tools used for assessing technical skill. (Louridas, Szasz, de Montbrun, Harris, & Grantcharov, In Press)

2.4.1 Traditional Platforms to Assess Technical Skill

Technical skill is evaluated during training and indirectly evaluated at the time of certification; however, the common methods used to evaluate technical skill have been criticized for being subjective, unstructured, invalid and unreliable. Traditional methods for evaluating technical skill during training included direct observation, surgical log books, in training evaluation reports (ITERs) and prescribed time of training.

2.4.1.1 Direct Observation

The apprenticeship model of training assumes that it is the responsibility of faculty to assess the technical proficiency of the trainee through direct observation. (Moorthy, Munz, Sarker, & Darzi, 2003; R. K. Reznick & MacRae, 2006) However, this subjective and unstructured assessment can suffer from significant bias; for example, Warf and colleagues demonstrated that senior trainees were more often perceived to be more “competent” than junior trainees despite a lack of difference in technical skill performance. (Warf, Donnelly, Schwartz, & Sloan, 1999)

2.4.1.2 Surgical Logs

Another common method to evaluate surgical performance is the maintenance of a surgical log book, in which the trainee keeps track of all procedures performed throughout his or her
training. (Grantcharov, Bardram, Funch-Jensen, & Rosenberg, 2002) While log books tell us about the quantity of cases, they do not inform us of the quality of the technical performance or how much of the procedure was completed independently, and thus are difficult to interpret and may not accurately reflect operative ability.

2.4.1.3 In Training Evaluation Reports

In training evaluations are used to assess several domains of competency, including technical skill, and are one of the main tools used to promote residents during their training and certify residents at the time of residency completion. Many factors can influence the perception of competence and thus the scores on an ITER. In training evaluation reports are subjective and, like direct observation, suffer from rater bias. For example, research suggests that the more familiar we are with an individual, the more positively we will evaluate him or her. (Warf et al., 1999) It has also been shown that ITERs fail to identify below average residents. (Feldman, Hagarty, Ghitulescu, Stanbridge, & Fried, 2004) For example, a resident who performs below the level of expected performance on an objective measure of technical skill will often still receive “satisfactory” on his or her ITER. Basing decision of promotion and certification on ITER data clearly needs to be reconsidered.

2.4.1.4 Prescribed Time

Surgical residencies have traditionally been time based, as it has been assumed that completing the prescribed “time” (e.g. five years for general surgery) will allow for the acquisition of competency in the respective surgical specialty. However, evidence suggests that the acquisition of technical skill varies among individuals, so time alone should not serve as a surrogate for the acquisition of technical competence.
2.4.1.5 Surrogates of Technical Skill

It is often assumed that if a surgical trainee is competent in the domains of knowledge and judgment, they are also competent in the domain of technical skill. However, measures of knowledge and judgment as a surrogate for technical ability is inappropriate, and it has been clearly shown that the results of reliable and valid measures of knowledge (e.g. the American Board of Surgery In-Training Exam, or ABSITE) do not correlate with technical skill or operative performance.(Scott et al., 2000)

2.4.1.6 Issues with Traditional Assessment Methods

All of these traditional methods of assessing technical skill lack the objectivity, reliability and validity that are needed to accurately assess skill and allow for the promotion and certification of residents. Recognizing the need for more rigorous assessment tools, several groups have developed objective assessments of technical skill, and studies looking at their reliability and validity have been published in the literature.(van Hove, Tuijthof, Verdaasdonk, Stassen, & Dankelman, 2010)

2.4.2 Objective Assessments of Technical Skill

The two most studied tools that have been developed to objectively assess technical skill are Procedure Based Assessments (PBA) and the Objective Structured Assessment of Technical Skill (OSATS).

2.4.2.1 Procedural Based Assessments (PBA)

Developed by Pitts and colleagues in Great Britain(Orthopedic curriculum and assessment project, 2006), the PBA is a workplace based assessment tool that has become the main method used in the UK to assess advanced technical skill in the operating room.(J. D. Beard, 2007)
Trainees are observed performing an index case and, at the end of the case, the PBA form is completed and returned to the trainee for immediate formative feedback. The PBA was originally designed as a formative assessment tool, with a strong emphasis on feedback and learning. However, research has shown that multiple evaluations of a resident performing the same operation can provide reliable summative data on performance. (J. Beard et al., 2009) (Pitts, Rowley, & Sher, 2005) The advantages of the PBA are that it does not require a testing infrastructure, since the assessment is done in the context of clinical cases. Furthermore, direct observation in the operating room is the “gold standard” for the assessment of surgical skill in terms of validity since it captures real live performance. However, there are several disadvantages when considering the PBA for the purpose of high stakes assessment such as certification. The major issue is the ethics of testing performance in the context of a real operation where residents would not be allowed to demonstrate poor performance. Secondly, the inherent variability of cases makes it impossible to standardize the operative task. Thirdly, in order to achieve reliable summative data, a large number of assessments for each type of operation would be needed, making the feasibility of using the PBA for certification difficult. Finally, the biases present in the in training evaluation report and subjective direct observation may also be present in the PBA since attending surgeons mark their own trainees. Thus, although the PBA is an excellent form of assessment for formative feedback, there are limitations in its use in high stakes assessment such as certification.

2.4.2.2 Objective Structured Assessment of Technical Skill (OSATS) Examination

The OSATS examination was developed in 1997 by Martin et al. from the University of Toronto. (Martin et al., 1997) Their goal was to create and evaluate a tool for assessing the technical skill of surgical trainees. This tool is a performance based examination made up of several time limited stations where candidates move from station to station performing a surgical task while being directly observed and evaluated by an expert surgeon. The observer uses both a task specific checklist and a global rating scale to assess performance. (Martin et al., 1997) Studies looking at the validity and reliability of OSATS have shown that it is acceptable for summative high stakes evaluation purposes, with evidence of sufficient reliability when eight
stations are used, and good evidence of validity. (Martin et al., 1997; R. K. Reznick & MacRae, 2006) It has also been shown to be feasible across a number of programs. (R. K. Reznick & MacRae, 2006) (van Hove et al., 2010) (Faulkner, Regehr, Martin, & Reznick, 1996)

The OSATS examination in its original description included the global rating scale, which included nine domains of technical skill, each providing a gestalt of technical ability. (Martin et al., 1997) The OSATS global rating scale has become one of the most commonly used assessment tools for assessing technical skill both inside and outside of the operating room. (Niitsu et al., 2013; Szasz et al., 2014) The global rating scale is not specific to a task and can be used to assess any technical procedure. The global rating scale is seen in Figure 2. Some have suggested that a checklist may reward thoroughness rather than competence, thus rewarding the action of a novice (who pays great attention to detail), versus recognizing the experience of an expert clinician (who may not be thorough but is very accurate). It has been argued that rigid checklists may actually have lower validity than global rating scales, and the literature supports this notion. Regehr et al. have demonstrated that global rating scales used by expert examiners during a performance based examination demonstrate better evidence of validity than checklist based evaluations. (Regehr, MacRae, Reznick, & Szalay, 1998) They suggested that when constructing a performance based examination, priority should be given to well-constructed global rating scales rather than task-specific checklists.

When considering a high stakes assessment, the OSATS simulation based assessment overcomes many of the drawbacks associated with the use of PBAs for certification. Since the OSATS examination takes place in the surgical skills laboratory, it allows standardization of the examination and also avoids the ethical issue of “testing” performance on real patients. For these reasons an OSATS modeled examination is ideal for high stakes assessment of technical skill and certification. However, one of the major limitations preventing the use of the OSATS examination in pass/ fail decisions, even acknowledged by one of its developers, is the lack of an established passing score. (R. K. Reznick & MacRae, 2006)
In order to be incorporated into a high stakes decision such as certification or resident promotion, an assessment tool requires the establishment of a point along the score scale that reflects an “adequate” performance. This is known as a passing score.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect for tissue</td>
<td>Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments</td>
<td>Careful handling of tissue but occasionally caused inadvertent damage</td>
<td>Consistently handled tissue appropriately with minimal damage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time and motion</td>
<td>Many unnecessary moves</td>
<td>Efficient time/motion but occasionally causes inadvertent damage</td>
<td>Economy of movements and maximum efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument handling</td>
<td>Repeatedly makes awkward moves with instruments</td>
<td>Competent use of instruments although occasionally appears stiff or awkward</td>
<td>Fluid moves with instruments and no awkwardness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of Instruments</td>
<td>Frequently asked for the wrong instrument or used an inappropriate instrument</td>
<td>Knew the names of most instruments and used appropriate instrument for the task</td>
<td>Obviously familiar with the instruments and their names</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of assistant</td>
<td>Consistently placed assistant poorly or failed to use assistant</td>
<td>Good use of assistant most of the time</td>
<td>Strategically used assistant to the best advantaged at all times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow or operation and forward planning</td>
<td>Frequently stopped operating or needed to discuss next move</td>
<td>Demonstrated ability for forward planning with steady progression of operative procedure</td>
<td>Obviously planned course of operation with effortless flow from one move to the next</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of specific procedure</td>
<td>Deficient knowledge. Needed specific instruction at most operative steps</td>
<td>Knew all important aspects of the operation</td>
<td>Demonstrated familiarity with all aspects of the operation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Objective Structured Assessment of Technical Skill (OSATS) global rating scale (Adapted from Martin et al., 1997)
A passing score is an essential component of a certification examination in order to differentiate between a competent and non-competent performance. Regardless of the validity or rigorousness of an assessment tool, decisions of competence are not possible without a passing score.

2.4.3 Current Performance/Procedure Based Assessments Used in High Stakes Decisions

Several performance based assessments within the domain of medicine have been developed and variably implemented into practice.

The Israeli Board of Anesthesia examination, the Objective Structured Clinical Examination (OSCE) of the Medical Council of Canada, the Fundamentals of Laparoscopic Surgery (FLS) examination, the European Board of Vascular Surgery technical skill examination and the Colorectal Objective Assessment of Technical Skill (COSATS) are all assessment tools that have been developed for the purpose of high stakes assessment within the domain of medicine.

2.4.3.1 The Israeli Board of Anesthesia Examination

Although anesthesia is not a surgical specialty, anesthesiologists are expected to possess the skills necessary to manage a wide range of acute intraoperative anesthetic events, many of which include technical skill in various types of procedures. Requisite technical skills for this specialty include, for example, endotracheal intubation, insertion of an arterial line, insertion of central lines and management of a difficult airway. Anesthesia as a specialty therefore requires that individuals develop these technical skills.
In the early 2000s the Israeli Board of Anesthesiology recognized the lack of performance evaluation at the time of board certification as well as the lack of performance assessment during the training process. At that time the Test Committee of the Board decided to develop, evaluate and subsequently implement an Objective Structured Clinical Examination (OSCE) type of examination as a component of the Israeli Board of Anesthesia certification process. (Berkenstadt et al., 2012; Berkenstadt, Ziv, Gafni, & Sidi, 2006a)

This initiative, although not derived from a surgical specialty, can be used as an example of the development, evaluation and implementation of a performance based assessment into a high stakes board certification examination.

The development and implementation of this examination took place in the early 2000s. The development of the examination involved several steps. Firstly, the content of the examination was selected through expert opinion, which identified the clinical skills that residents would be expected to handle competently at the end of their training. A Delphi methodology was used to identify tasks that would represent those skills, and scenarios assessing those tasks were created. Simulated models were then developed to assess those chosen tasks. The models were then piloted by junior attending staff prior to implementation. The tasks included scenarios such as management of a trauma casualty, hypertension after induction of a general anesthetic, regional anesthesia land marking on a standardized patient, management of convulsions, and adjustment of ventilator setting in response to an arterial blood gas.

A passing score based on checklist scores was arbitrarily set at 70%. An overall holistic assessment scale was also used. In 2006, Berkenstadt et al. published results of the first 114 trainees taking the examination over the two year pilot study. (Berkenstadt, Ziv, Gafni, & Sidi, 2006b) Inter-station reliability was 0.35 – 0.45 and acknowledged to be low. This was felt to be due to the limited number of stations (case specificity), and the limited number of candidates (n=17 per year). The results demonstrated a weak correlation with the written and oral examination, suggesting that they are measuring a different construct.
The examination became a mandatory component of the Israeli Board of Anesthesia examination and continues to be a component of their certification process. One of the weaknesses of this examination is the arbitrarily selected passing score, and lack of a rigorous standard setting methodology. Standard setting methodologies have limitedly been used in performance based assessments in medicine.

2.4.3.2 The Medical Council of Canada (MCC) Objective Structured Clinical Examination (OSCE)

The Medical Council of Canada (MCC) has a legislated national mandate to ensure that Canadian medical doctors meet the same demanding, consistent standards across the country. To ensure this, the MCC requires that individuals complete two examinations, the Medical Council of Canada Qualifying Examination (MCCQE) Part I and Part II. Part I (MCCQE Part I) assesses the competence of candidates who have obtained their medical degree or entry into supervised clinical practice in postgraduate training programs and consists of a written examination. The Medical Council of Canada Qualifying Examination Part II consists of a multi-station Objective Structured Assessment of Clinical Skill (OSCE) and is required for medical licensure in Canada prior to entry into independent clinical practice. The OSCE for the Licentiate of the Medical Council of Canada (LMCC) Part II is one of the best examples of the incorporation of a performance based assessment into medical certification.

In the late 1980s the MCC recognized that many competencies expected of licentiate candidates (e.g. history taking, physical examination, communication skills) were not being evaluated in the existing examination process. The OSCE, which evaluates the clinical skills not addressed in written examinations, was then piloted and subsequently implemented into the Medical Council of Canada Qualifying Examination.
The MCC conducted several pilot studies looking at the OSCE and the logistics of incorporating this performance-based examination into high-stakes, large-scale testing for the purpose of licensure. There were approximately 2,200 individuals who would be eligible to take this examination every year.

The original OSCE pilot project in 1991 was administered to 240 paid volunteer candidates in two Canadian cities. This initial study demonstrated that the examination maintained its psychometric properties across different sites, suggesting that large-scale administration of the examination would be feasible. (R. Reznick et al., 1992) A second pilot was conducted in 1992 for the purposes of further exam development and psychometric evaluation testing of a larger cohort of candidates (n=401). (R. K. Reznick et al., 1993) Following the results of these two pilot studies and with the push from regulatory bodies and the Medical Council of Canada, the OSCE became a mandatory component of Canadian licensure in 1994. The Medical Council of Canada and the provincial regulatory bodies were eager to incorporate this novel assessment into licensure, recognizing that the competencies evaluated were essential to clinical practice and were previously not assessed. Based on the results of these initial studies and through an act of parliament, a bi-law was passed and the OSCE became a mandatory component of licensing in 1994. Thus, as of 1994, completion of the MCCQE Part II became a necessary step for medical licensure in Canada.

2.4.3.3 The Fundamentals of Laparoscopic Surgery (FLS) Examination

The technical skill component of the Fundamentals of Laparoscopic Surgery (FLS) program is based on the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS). (Vassiliou et al., 2006) The MISTELS is an assessment of basic laparoscopic skills, including peg transfer, cutting, placement of a ligating loop, and intra and extracorporeal knot tying. This tool has been extensively studied and validity evidence has been demonstrated in the literature, with the ability to discriminate between expert and novice laparoscopic surgeons. (Fraser et al., 2003) The MISTELS was the first technical skill assessment to use a methodologically sound approach to setting a pass/fail score, and has thus established the ability
to assign a pass/fail standing. The FLS program is not a Board certification examination per se, however, since 2009 it has become a mandatory prerequisite for all candidates taking the American Board of Surgery certification exam. (Surgery, 2012) The major disadvantage of the FLS examination is that it assesses very basic laparoscopic skills, such as knot tying and PEG transfer, limiting its ability to assess advanced technical skill for certification.

2.4.3.4 The European Board of Vascular Surgery Examination

Back in the early 2000s, the European Board of Vascular Surgery (EBVS) recognized the lack of formal assessment of technical skill at the time of completion of training, and the Board decided to develop a technical skill exam for the certification of vascular trainees. A technical skill examination consisting of three vascular surgical tasks on bench-top models was developed and evaluated. Published results of this pilot study demonstrated initial evidence of construct validity, inter-rater reliability and good internal consistency of the exam. (Pandey, Wolfe, Liapis, & Bergqvist, 2006) Construct validity was established by comparing the performance of experts (surgeons in practice for over five years) to trainees, and finding a statistically significant difference in performance between the two groups. Since 2004, this technical skill examination has been incorporated into the Board certification process. One of the limitations of this assessment tool is the lack of a rigorous methodology for setting a standard for the exam, with an arbitrarily set passing score of 75%. The Board has not published any further data on the ongoing development and evaluation of this tool.

2.4.3.5 The Colorectal Objective Structured Assessment of Technical Skill (COSATS) Examination

The recognition of a lack of objective assessment of technical skill at the time of surgical certification has also been acknowledged by the American Society of Colon and Rectal Surgeons (ASCRS). In response to this gap in assessment, the ASCRS, working in conjunction with the American Board of Colon and Rectal Surgery (ABCRS), has developed an objective assessment
of technical skill for the purpose of board certification of colorectal residents, the Colorectal Objective Structured Assessment of Technical Skill (COSATS). The COSATS exam is a bell-ringer exam that takes place in the skills laboratory. The examination is made up of bench-top simulated models made from synthetic material and/or animal tissue. One station consisted of a virtual reality platform to assess colonoscopic skills. Examples of the simulated COSATS models are outlined in Figure 3.

The COSATS mirrors the structure of the OSATS exam whereby candidates rotate through eight stations performing a technical task specific to the practice of colon and rectal surgery. The initial pilot study compared general surgery residents to colorectal residents who were all at the end of their training, and found that the COSATS exam had the ability to differentiate between these two groups. (S. L. de Montbrun et al., 2013) This built initial evidence of construct validity. Furthermore, the COSATS demonstrated good inter-station reliability (Cronbach’s alpha 0.69) for a homogeneous group of examinees.

With the success of this initial pilot, a second study was conducted to further evaluate the reliability, validity and feasibility of the COSATS as well as to set a passing score. This second pilot study (unpublished data) assessed the performance of 37 volunteers who were
Figure 3. Simulated COSATS models a. ileal pouch anal anastomosis b. presacral bleeding c. coloanal anastomosis d. perianal fistula e. loop ileostomy f. virtual reality colonoscopy
concurrently taking their ABCRS oral board exam. Results demonstrated that the COSATS could be administered in conjunction with the ABCRS oral exam, demonstrating its feasibility. The study also used a borderline group methodology to set standards of performance and assign a pass/fail standing to the candidates. The failure rate of the COSATS was 8%, which was consistent with the failure rate of the oral (8%) and written (8%) portions of the examination. However, individuals who failed the COSATS were not the same individuals who failed the oral or written portions of the exam, suggesting that the COSATS does measure a different construct (i.e. technical skill) than the written and oral exam, and thus adds greater validity to the exam as a whole. These results also suggested that some individuals who are being certified by the Board as competent colorectal surgeons (having passed the written and oral board examination) are entering practice despite having demonstrated a lack of technical ability on this novel technical skill assessment tool.

2.5 The Current Board Certification Process

The main goal of certification is to ensure that a candidate is competent in all facets required of his or her profession. Moreover, the certification process distinguishes between individuals who have adequate knowledge, skill and ability and those who do not. (D'Costa, 1986)

It is this assessment of clinical competence and the certification process that is intended to ensure that individuals entering independent practice have the necessary knowledge, judgment and skill to provide safe patient care. (Petrusa, 2009)

Certification exams and other high stakes assessments where important decisions are being made require that validation studies support the interpretive argument that performance on the test translates into or reflects performance in the real clinical setting. (M. Kane, 2013) Rigorous scientific investigation is required to provide validity evidence for the interpretation of test scores.
Presently, within the United States and Canada, surgical board certification includes the evaluation of several domains of competence. However, the assessment of technical skill is not directly assessed or evaluated at the time of certification. This gap in assessment fails to address a surgeon’s technical skill.

2.5.1 Canadian Surgical Board Certification

At the end of five years of general surgery training, Canadian general surgery residents must meet certain eligibility criteria in order to sit the Royal College of Physicians and Surgeons of Canada examination. Once eligibility is established, individuals can take their certification examination. Successful candidates are recognized nationally as having the credentials to independently practice general surgery. (The Royal College of Physicians and Surgeons of Canada, 2014a)

The Canadian general surgery board examination includes several components. Firstly, there is a written component, made up of two 3-hour exams, consisting exclusively of multiple choice questions. These questions are focused on general surgery and clinically applied basic science. Secondly, the examination has an oral and clinical component. This part of the exam is made up of eight stations. At each station, examiners ask candidates questions regarding the management of general surgical issues. This examination is two hours in duration. (The Royal College of Physicians and Surgeons of Canada, 2014d)

At the present time, there is no direct, formal, objective assessment of technical skill. The component of the certification process that addresses technical skill is the Final In Training Evaluation Report (FITER)/Comprehensive Competency Report (CCR). This is a written document that is required to be filled out by the general surgery program director, indicating that the resident has met the necessary requirements of the specialty and is eligible to write the Royal College examination in general surgery. (The Royal College of Physician and Surgeons of Canada, 2010) This form requires that the program director respond with yes or no to the following statement:
“In the view of the Residency Program Committee, this resident has acquired the competencies of the specialty/subspecialty as prescribed in the Objectives of Training and is competent to practice as a specialist.” (The Royal College of Physician and Surgeons of Canada, 2010)

Data informing the response to this question on the FITER comes from several sources, including clinical observations from faculty (e.g. In Training Evaluation Reports – ITERs). This is the only component of the certification exam that directly addresses the issue of technical competence. From the literature on ITERs, it is known that the ITER may not be a reliable indicator of performance, and thus having an assessment tool to directly assess technical skill at the time of certification that is both reliable and valid would be ideal.

Furthermore, most subspecialty surgical board certification also currently consists of the assessment on knowledge and judgment without the formal, direct assessment of technical skill. (The Royal College of Physicians and Surgeons of Canada, 2014e)

2.5.2 American Surgical Board Certification

The board certification process in the United States for general surgery certification includes two components. The first is the General Surgery Qualifying Examination. This component of the examination is structured as a multiple choice exam, made up of around 300 questions. The domains of competence that are evaluated include knowledge of general surgical principles and application of scientific knowledge to general surgical practice. The exam takes place over one day, and is eight hours in duration. (The American Board of Surgery, 2015)

The second and final step of certification is the General Surgery Certifying Examination, which is an oral examination focused on the evaluation of clinical skills, focusing on diagnosis and management of common surgical problems, and evaluating judgment, reasoning and problem solving. (The American Board of Surgery) Although questions may focus on the description of the technical details of an operation, as in the Canadian certification process, technical skill is not directly assessed.
As in Canada, the board certification for subspecialties in surgery in the United States, for the most part, also consists of a written and oral component. (American Board of Colon and Rectal Surgery, 2012)

Surgical boards and regulatory bodies have a keen interest in the evaluation of competence of surgical trainees, to ensure that those individuals entering independent practice are safe. However, the current system of certification is missing the assessment of an essential and hallmark component of competence for surgical specialties. Interestingly, the intraoperative performance of surgeons prior to the 1950s was evaluated by the American Board of Surgery through observation. However, and not surprisingly, the practicality and feasibility of this approach precluded its ongoing use as a form of assessment. (Bell, 2009)

The assessment process is lacking the evaluation of technical skill of surgical trainees. It is time to develop, evaluate and implement objective assessments of technical skill as a component of certification.

In this regard, one of the first important steps is having a working definition of “competence” that can be applied to the interpretation of assessment tools.

2.6 Competence

2.6.1 Definition

Use of the word “competence” within the domain of surgery has been inconsistent, with the terms technical competence, surgical competence and operative competence being referenced in the literature.

Satava et al. have defined competence in technical skill as an individual who is able to answer all of the requirements; be suitable, fit and adequate. (Satava, Gallagher, & Pellegrini, 2003) The goal of a high-stakes exam such as certification is to identify individuals who have reached a
certain predefined level of competence. This threshold needs to be fair to both the public to protect their safety, as well as to the individual who has invested time and effort into their training.

A more recent systematic review aligns with this definition and has suggested that the term technical competence has most often been described in the literature as being an individual who possesses a minimum standard of performance to provide safe and independent surgical care. (Szasz et al., 2014)

Szasz et al. suggest that surgical competence represents a collection of skills, knowledge and judgment in both technical and non-technical skill, while operative competence includes surgical experience and the ability to apply skills to a broader domain. (Szasz et al., 2014)

Technical competence as representing an individual who is suitable, fit and possessing a minimum standard of performance to provide safe care, will be the definition used throughout the following studies.

2.6.2 The Royal College of Physicians and Surgeons CanMEDS Framework

Surgical competence is certainly made up of several domains including but not limited to knowledge, judgment and technical skill. A broader scope of competency is depicted in the Royal College of Physicians and Surgeons of Canada CanMEDS model, which outlines a framework for understanding the abilities required of a competent and safe physician. (The Royal College of Physicians and Surgeons of Canada, 2014b) Seven key roles are outlined in this model. This comprehensive framework has been used around the globe and is depicted in Figure 4.
These competencies require evaluation throughout and upon completion of training. Certainly for surgeons an essential domain of competence, surgical skill is a domain of competence that would fall within the medical expert category, and is a key competence that requires assessment.

**Figure 4.** The Royal College of Physicians and Surgeons of Canada CanMEDS framework (Adapted from [http://www.royalcollege.ca/portal/page/portal/rc/canmeds/framework](http://www.royalcollege.ca/portal/page/portal/rc/canmeds/framework))
2.6.3 Accreditation Council for Graduate Medical Education

The Accreditation Council for Graduate Medical Education (ACGME), in the 1990s defined the necessary components of surgical competence. (Mery et al., 2008) The six domains outlined were 1) medical knowledge 2) patient care 3) professionalism 4) communication skills 5) practiced-based learning and 6) systems-based learning.

For surgeons, technical skill falls under the domain of patient care. The ACGME, continues to work with specialty groups to develop methods to assess these six domains of competence. A key component of the development of milestones involves the creation of valid and reliable assessment tools to evaluate technical skill. (Cogbill, Malangoni, Potts, & Valentine, 2014)

Developing tools to assess technical skill is essential as surgical regulatory bodies continue to require the documentation of the achievement of competence. Milestone projects also struggle with the lack of reliable and validated tools to assess these competencies.

For an assessment tool to be used in high stakes decisions, such as certification or resident promotion, it needs to be continually validated. These validation studies over time can build enough evidence to allow the tools to be used in high stakes decisions.

2.7 Theories of Validity

Validity is not only the most important characteristic but also the *sine qua non* of assessment data. Validity refers to the evidence presented to support or refute the interpretations of test results or scores (Messick, 1989), and validation studies look to collect evidence to support the interpretation of assessment results. An assessment is tool is never said to be “valid” or “invalid”, rather it is the interpretations of the assigned scores that are evaluated in a validation study. (Downing & Haladyna, 2009) Validity evidence is this both time and case specific; validity evidence generated from a study cannot automatically be applied to a different time
period or different population. For example, a test that is developed to assess advanced technical skill may allow for the valid interpretation of scores when applied to senior surgical trainees; however, when the same test is applied to medical students it is unlikely that valid judgments of their ability can be made because the task is simply too difficult. Validity can therefore be understood as a property of inferences made from scores rather than the property of the instrument itself, with validity evidence needing to be established for each intended interpretation.(Cook & Beckman, 2006) Furthermore, the level of validity evidence that is needed to support the use of an assessment tool depends on the purpose and goal of the assessment. For example, a high stakes decision, such as certification or graduation, would require a high level of validity evidence because the consequences of the decisions are significant. A high stakes assessment requires the ongoing collection of validity evidence from multiple sources because of the significance of the decision; furthermore, summative assessments require a wider breadth and more rigorous validity evidence than formative assessments.

Over the past several decades psychometricians and measurement experts have re-conceptualized validity, moving from a fragmented traditional model to a unified contemporary model. This shift in conceptualization dates back a few decades, however, the domain of surgical education has significantly lagged in the application of the contemporary model of validity and the literature has predominantly described assessment tools using the traditional model.

2.7.1 Traditional Model of Validity

Validity was previously divided into three types: content validity, criterion validity and construct validity.

Content validity referred to the degree to which the test adequately assessed the content domain of interest.(Lawshe, 1975) So, for example, an assessment aimed to assess surgical knowledge should include items that would be considered important and essential to surgical knowledge and items that would be a part of the surgical curriculum.
Criterion validity refers to the extent to which a measure is correlated with a second measure. If this comparison is done with a measure taken at the same time, it is referred to as concurrent validity. If this comparison is done with a measure taken at a future point in time, it is referred to as predictive validity. (Cronbach, 1990)

Construct validity was described as the extent to which the assessment measured the construct it was intended to measure. (Moorthy et al., 2003) This type of validity evidence abounds in the surgical literature. A recent systematic review of validity evidence used in simulation based assessment found that the single most common source of validity evidence studied was from comparing scores between training or experience levels. (Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014) These studies typically consider that individuals with more training or experience, who we assume possess more of the construct, will outperform individuals with less experience or training, who we assume possess less of that same construct.

Face validity is a component of validity that has also been popular in medical education and was described as representing the degree to which the assessment “looked” valid. This component of validity has been highly scrutinized by measurement experts and is not endorsed by contemporary psychometricians. (Downing, 2006) Face validity is not regarded as a legitimate source of validity evidence and should be abandoned from any validation studies.

This classic classification of three different types of validity has given way to contemporary unified models of validity, whereby all validity evidence is considered construct validity. Two robust contemporary approaches to validity theory exist. Both focus on how validity evidence supports the interpretation of assessment scores. While many of the same concepts of the traditional validity model are still relevant in the contemporary models, they have been reclassified and conceptualized in these new unified frameworks.
2.7.2 Contemporary Validity Theory

While the field of assessment validity has evolved significantly over the past several decades, with a contemporary model proposed in 1989 by Messick (Messick, 1989), the surgical education literature continues to frequently use the fragmented traditional framework. (Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013; Ghaderi et al., 2015) In 2010, Korndorffer et al. evaluated the surgical education literature for the adoption of contemporary validity frameworks in the development and evaluation of surgical assessment tools and found that 100% of studies in their analysis used the traditional framework of validity, with 75% using construct validity, 38% using face validity and 11% using content validity. (Korndorffer, Kasten, & Downing, 2010) Two current systematic reviews investigating current available methods of technical skill assessment also report on validity using the traditional validity framework. (Ahmed, Miskovic, Darzi, Athanasiou, & Hanna, 2011; van Hove et al., 2010)

The present PhD studies are grounded in the current contemporary validity frameworks, systematically collecting evidence to test the hypothesis that scores are validly interpreted for their use. The present studies will help to continue to introduce and apply contemporary validity frameworks to the surgical assessment literature and remain current within the domain of educational measurement.

2.7.2.1 Messick’s Validity Model

In feeling that the traditional model of validity was fragmented and incomplete, Messick describes a framework of validity in which five distinct sources of validity evidence are sought in order to build a validity argument. (Messick, 1989) A validity hypothesis is tested by examining the evidence to refute or support the interpretation of test scores. The amount of validity evidence required again depends on the goal and purpose of the evaluation, with validation studies aiming to gather evidence from various sources.
Messick’s five sources of validity evidence are: 1) content 2) response process 3) internal structure 4) relationship to other variables and 5) consequences. (Messick, 1989) Definitions and examples of Messick’s sources of validity are outlined in Table 3.
Table 3. Messick’s five major sources of validity evidence

<table>
<thead>
<tr>
<th>Source of Validity Evidence</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Relationship between test content and the construct of interest</td>
<td>Examination blueprint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Representativeness of items to domain</td>
</tr>
<tr>
<td>Response Process</td>
<td>Data integrity, such that all sources of error are controlled or eliminated to the maximum extent possible</td>
<td>Rater training</td>
</tr>
<tr>
<td></td>
<td>Analyses of responses (actions, strategies, thought process) of individual respondents or observers</td>
<td>Student format familiarity</td>
</tr>
<tr>
<td>Internal Structure</td>
<td>The statistical or psychometric characteristics of the examination</td>
<td>Score scale reliability</td>
</tr>
<tr>
<td></td>
<td>Internal consistency, reliability, generalizability</td>
<td>Psychometric model</td>
</tr>
<tr>
<td></td>
<td>Degree to which individual items within an instrument fit the underlying construct</td>
<td>Item factor analysis</td>
</tr>
<tr>
<td>Relationship to Other Variables</td>
<td>Relationship of scores to other variables relevant to the construct being measured</td>
<td>Convergent correlation with tests measuring the same construct</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Divergent relationship with tests measuring different constructs</td>
</tr>
<tr>
<td>Consequences</td>
<td>Impact on examinees from the assessment scores, decisions and outcomes</td>
<td>Pass/fail reliability</td>
</tr>
<tr>
<td></td>
<td>Impact on teaching and learning</td>
<td>Impact of test scores on individuals</td>
</tr>
<tr>
<td></td>
<td>Impact on faculty, patients and society</td>
<td>Standard setting methodologies</td>
</tr>
</tbody>
</table>

Adapted from Downing (2003)(Downing, 2003) and Ghaderi et al. (2014)(Ghaderi et al., 2015)
2.7.2.2 Kane’s Validity Model

While the surgical education assessment literature is slowly starting to become current in validity theory by framing validation studies using Messick’s contemporary approach, there has been a paucity in applying Kane’s model despite this framework being very applicable to the development and evaluation of medical assessment. In a recent systematic review, Kane’s framework was used to build a validity argument for the Objective Structured Assessment of Technical Skill (OSATS). (Hatala, Cook, Brydges, & Hawkins, 2015) Prior to this, Kane’s model had not been used for technical skill assessment.

Kane’s argument-based approach to validation uses a framework to address the inferences made from test scores. (M. Kane, 2013) The first step in Kane’s model is to outline the intended use of the assessment scores and then evaluate evidence based on four validity inferences. The four inferences in Kane’s validity argument are: 1) scoring 2) generalization 3) extrapolation and 4) implications/decisions. (M. Kane, 2006)

Building evidence for each inference will progressively lead towards translating an observation on an assessment, to a measured score, inferring performance on the test and in the real world setting and extrapolating these results to specific actions. (Hatala et al., 2015) Kane’s validity argument with four inferences is outlined in Table 4.
Table 4. Definition of the four inferences in Kane’s validity argument

<table>
<thead>
<tr>
<th>1) Scoring</th>
<th>2) Generalization</th>
<th>3) Extrapolation</th>
<th>4) Implications/decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule is appropriate</td>
<td>Sample is representative of universe of possible observations</td>
<td>Observed score is related to the target score</td>
<td>Implications are appropriate</td>
</tr>
<tr>
<td>Rule is applied as specified</td>
<td>Sample is large enough to control for random error</td>
<td>No systematic errors likely to undermine the extrapolation</td>
<td>Properties of scores support the implications associated with the label</td>
</tr>
<tr>
<td>Scoring is free of bias</td>
<td>Data fit the scaling model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Adapted from Kane 2006 pg. 34)(M. Kane, 2006)

It is imperative that educationists in surgical skills assessment stay current with theories of validity, applying them to ongoing study in order to rigorously develop and appropriately evaluate tools. This is of utmost importance when developing high stakes summative assessments, where decisions based on inferences from test scores have significant impact on both the individual and society.

2.8 Developing an Examination Blueprint

2.8.1 Content Evidence – Messick’s Framework of Validity

The foundation of validity evidence for an assessment lies in the development of its content. Content evidence in Messick’s framework of validity refers to the specifications of the construct being assessed(Messick, 1989), and considers the relevance and representativeness of tasks included in the assessment.
Building content validity evidence would include the description of the process used to ensure that the content of the assessment reflects the construct being assessed.

A construct has been defined as:

“…a theoretical entity – something that we believe exists and which can be described, but which may not be amenable to direct measurement.” (Andreatta & Gruppen, 2009)

A construct within surgical training would be technical skill. This could then be more clearly defined, such as technical skill for junior residents, or technical skill and procedures specific to graduating general surgery residents. When developing exam content, the construct of interest needs to be clearly defined and represented in the assessment.

Exam blueprinting is the process of building exam content that reflects the objectives of the curriculum or training program. (Newble, Jolly, & Wakeford, 1994)

The use of expert opinion has been one of the most commonly used techniques in developing the content of technical skill assessment tools. A recent review suggests that most technical skill assessment tools have used expert opinion to create exam content, including information from literature reviews and textbooks. (Ghaderi et al., 2015)

More rigorous consensus methodologies are appropriate methods to build a valid blueprint for an assessment tool. Furthermore, using a methodologically sound approach in developing the content of a high stakes examination is essential in helping to ensure the legal defensibility of the test content and interpretation of the test scores. (Castle, 2002) This is an essential first step in developing a high stakes assessment.
2.8.2 Delphi Methodology

Delphi methodology is a process for achieving expert consensus and has been used in medical education for the development of curricula (Palter, Graafland, Schijven, & Grantcharov, 2012; Palter, MacRae, & Grantcharov, 2011; Zevin, Levy, Satava, & Grantcharov, 2012) and training frameworks.(Bonrath, Dedy, Zevin, & Grantcharov, 2014; Szasz, Louridas, de Montbrun, harris, & Grantcharov, 2016) The Delphi technique uses statistical aggregation of expert opinion in an anonymous fashion to gain consensus over two or more iterations.(Goodman, 1987; Graham, Regehr, & Wright, 2003) This consensus methodology applies well to the development of a technical skill assessment blueprint, and was used in the development of a blueprint for a technical skill examination for Canadian general surgery residents, described in chapter 3.

2.9 Standard Setting

If the purpose of a high stakes, summative assessment is to define competency, it is imperative that valid, defensible cut-off scores be determined in order to correctly delineate between a competent candidate and a non-competent candidate.

Addressing this issue of setting a passing score and exploring the outcomes of a set score builds evidence of “consequences” within Messick’s framework of validity. This is an area of validity that has been understudied within the domain of technical skill assessment.

2.9.1 Definition of a Passing Score and Performance Standard

Standard setting methodologies are processes that are used to set a cut point or passing scores for an assessment tool. This specific point reflects the minimum level of competence deemed necessary for the content of the exam.(J. J. Norcini, 2003)

In the setting of certification or licensure exams, the consequences of a false negative or false positive result can have significant implications. A false positive may lead to patient harm by
passing candidates who should have failed the exam, while a false negative will lead to candidate injustice by failing someone who should have passed the exam. (Downing, 2003)

This score is an essential component of high stakes assessments as it allows for the differentiation between a competent and non-competent performance. While standard setting methodologies have a long history of use within the domain of medicine for written examinations, their use in performance based assessments in surgical skill assessment has been extremely limited. This is an area of education scholarship that requires further investigation.

While a *score* is the numerical representation of a candidates’ performance on the score scale, a *performance standard* represents the conceptualization of the performance that meets the expected level of competence for the purpose of the exam. (J. Norcini & Guille, 2002)

The score represents the numerical point along the score scale that conceptualizes the competent performance.

### 2.9.2 Setting a Passing Score

The purpose of standard setting is to establish a score that differentiates between those who perform well enough and those who do not. (J. J. Norcini, 2003) Processes used to set a passing score are referred to as standard setting methodologies. Much of the medical literature on standard setting has been in the domain of cognitive knowledge, with less known about setting standards for performance based assessments. (Schindler, Corcoran, & DaRosa, 2007) The largest body of literature looking at standard setting in the domain of performance based assessments comes from the Objective Structured Clinical Examination (OSCE). (Burrows, Bingham, & Brailovsky, 1999; Clauser & Clyman, 1994; Cusimano & Rothman, 2004; Dauphinee, Blackmore, Smee, Rothman, & Reznick, 1997; Jalili, Hejri, & Norcini, 2011; Kaufman, Mann, Muijtens, & van der Vleuten, 2000; Kramer et al., 2003; Livingston & Zieky, 1989; Medical Council of Canada, 2013b; Morrison, McNally, Wylie, McFaul, & Thompson,
Several standard setting methods have been described in the literature for performance based assessment. Common amongst all methods is the need to rely on the value judgments of experts in setting a passing score, resulting in a somewhat arbitrary decision. (Schindler et al., 2007) However, despite being arbitrary, a standard must still be credible. Norcini and Guille have described the criteria needed for a standard to be credible: 1) it must be set by an appropriate number and type of judges 2) it must use an appropriate methodology and 3) it must produces reasonable outcomes. (Norman, Vleuten, & Newble, 2002)

2.9.3 Absolute versus Criterion Referenced Standards

Standard setting procedures can be broadly categorized into either norm-referenced (relative) approaches or criterion referenced (absolute) approaches. Norm referencing describes an individual’s performance relative to the group. For example, a cut point is set to separate the top 20% from the bottom 80%. The purpose of this approach is to rank candidates, and is most appropriate when wishing to identify a certain percentage of examinees, for example for admission into medical school. In contrast, criterion referencing determines a standard score, for example 80% correct responses, that defines a competent level of performance. (Downing, Tekian, & Yudkowsky, 2006) The purpose of this approach is to determine if an individual has mastered a minimum level of performance. In theory, this approach can allow all candidates to pass if they achieve the criterion reference, and the pass/fail decision for any individual is irrespective of the group’s performance.

There is no “gold standard” method for establishing a passing score on a performance based competency exam, however, it has been suggested (Newble, 2004) that a criterion based approach is the most appropriate since it establishes a student’s level of competence based on a specified standard of achievement rather than on the performance of other members in the group. (Barman, 2008) Furthermore, absolute methods of standard setting are more appropriate in the setting of
high stakes examinations, where decisions of competence and licensure are being made. (McIlhenny & Orr, 2002; J. J. Norcini, 2003)

2.9.4 Test-Centered versus Participant-Centered Standard Setting

Test centered (also known as item centered) methods rely on expert judges to review the content of each test item and establish the performance of a borderline candidate on each item. (Swanson, Dillon, & Ross, 1990) The passing score is thus based on judgments about test questions. The overall passing score is the result of the aggregation of scores on each item, with the judges’ attention on the content of the test. (Livingston & Ziesky, 1982) An example of a test centered methodology is the Angoff method, which has a long history in standard setting in medicine for written examinations and has been used in establishing passing scores in performance based assessments in medical education. (Cusimano, 1996; Jelovsek et al., 2010; Kaufman et al., 2000; Kramer et al., 2003; Newble, 2004; Schoonheim-Klein et al., 2009) The Angoff method is the most commonly used methodology for setting standards on licensing and certification exams (M. T. Kane, Crooks, & Cohen, 1999) and is the most researched of all standard setting methods. (Cizek & Bunch, 2007)

The second approach is the participant centered (also known as examinee centered) methods. With these methodologies a passing score is based on judgments of examinee performance, rather than on the exam content. (Wilkinson et al., 2001) These methods rely on examiner observations and ratings that can be captured during the actual assessment, making them efficient and simple to implement. It makes intuitive sense that these approaches would apply well to performance based assessments, as the examinees performance can be used to set the passing scores. (Liu & Liu, 2008)

Three examinee based methodologies that can be easily applied to performance based assessments in surgery are 1) the borderline group methodology 2) the borderline regression methodology and 3) the contrasting groups method. In addition to being easily applied to performance based assessments in surgery, these methodologies are all conceptually relatively
straightforward and require basic statistical analysis, making them available to educational administrators.

2.9.5 Borderline Group Standard Setting Methodology

The borderline group method is based on the idea that the passing score be set at the level of a “borderline candidate” (Livingston & Ziesky, 1982). The borderline candidate is a key concept with most of the criterion referenced standard setting techniques (Downing et al., 2006). Angoff was the first to introduce the concept of the borderline candidate.

The borderline individual is someone who has an exact 50:50 probability of passing or failing the exam; he or she is the marginal student who on one day knows just enough to pass the exam, but on another day does not (i.e. an individual who is neither adequate or inadequate) (Downing et al., 2006; Livingston & Ziesky, 1982). This individual sits right on the “border” of the passing score.

The borderline group method attempts to set the passing score at the level of the borderline candidate. This involves expert judges identifying a group of examinees from the cohort that would be classified as “borderline”. This is done by having examiners give examinees both their score for the station as well as an overall categorical score of “borderline” (Newble, 2004).

For example, for a technical skill exam using the OSATS global rating scale, an examiner would provide each candidate with two scores 1) their global rating scale score and 2) an ‘overall performance’ evaluation, which would be based on a scale including the description of a borderline candidate. The passing score for a single station would be calculated as the mean score from all of the examinees classified as “borderline”. This methodology is schematically represented in Figure 5. The overall exam score would then be calculated as the sum of the passing scores for all stations. (Newble, 2004)
One of the main disadvantages of the borderline group method is its use in small scale examinations where there is a risk that the cut score is based on the performance of a relatively small number of individuals. (Wood et al., 2006) Within the domain of surgical education it is not uncommon to have a relatively small sample size, limiting the use of this methodology.

2.9.6 Borderline Regression Standard Setting Methodology

The borderline regression method has the ability to overcome the limitation of having a small sample size by using a linear regression model to regress all of the examinees scores onto their “overall performance” score. This overall performance scale would include, for example, the description of candidate performance from incompetent to expert competent. An example of an “overall performance scale” is seen in Figure 7.

Figure 5. Schematic representation of the borderline group standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015)
The point along the “overall performance” scale that determines where we expect this borderline candidate to fall – for example, at the descriptor of “competent for independent practice”, as an example a score of 3 on the overall performance scale seen in Figure 7, described as a borderline performance – can then be inserted into the linear equation to determine the predicted passing score for that station. The schematic representation of the borderline regression methodology is seen in Figure 6.

The passing score for each station is established and the overall exam score would be the sum of all station passing scores.

**Figure 6.** Schematic representation of the borderline regression standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015)
<table>
<thead>
<tr>
<th>OVERALL PERFORMANCE DESCRIPTORS</th>
<th>UNSATISFACTORY</th>
<th>NON-COMPETENT</th>
<th>BORDERLINE</th>
<th>CLEARLY COMPETENT</th>
<th>EXPERT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>This candidate clearly requires more training and was grossly inadequate in performing this task.</td>
<td>This candidate demonstrated basic technical ability, but inadequately performed this task.</td>
<td>This candidate is adequately competent to complete this task safely and independently.</td>
<td>This candidate is clearly competent to complete this task independently and safely. There is room for this individual to develop further expertise in technical ability.</td>
<td>This candidate demonstrated exceptional performance of this procedure.</td>
<td></td>
</tr>
<tr>
<td>Would not be comfortable with this candidate performing this procedure independently in training.</td>
<td>Would not be comfortable with this candidate performing this procedure independently in practice.</td>
<td>The candidate could perform this procedure independently in practice.</td>
<td>Fully comfortable with this candidate performing this procedure independently in practice.</td>
<td>This candidate is technically sophisticated in performing this procedure. I am fully comfortable with this candidate performing this procedure independently and safely in practice.</td>
<td></td>
</tr>
<tr>
<td>Unsafe to operate independently. Below the level of a colorectal resident.</td>
<td>This candidate would require significant supervision. At the level of a colorectal resident who still requires some training.</td>
<td>This candidate’s performance would fall RIGHT ON THE BORDER of the “pass” score.</td>
<td>This performance is ABOVE the passing score.</td>
<td>Exemplary technical skill, with a performance WELL ABOVE the passing score.</td>
<td></td>
</tr>
<tr>
<td>This performance is a CLEAR “fail”, and WELL BELOW the passing score.</td>
<td>This performance is BELOW the passing score.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7.** Example of an overall performance scale (Adapted from de Montbrun S., Satterthwaite L., Roberts P., MacRae H. 2016)
2.9.7 Contrasting Groups Standard Setting Methodology

The contrasting groups method is based on the idea that test takers can be divided into two groups, those who clearly “pass” and those who clearly “fail” the exam, or a “clearly competent” group and a “clearly non-competent” group. (Livingston & Ziesky, 1982) Once these two groups have been identified, their data is used to establish the passing score. Histograms of the scores from both the passing group and the failing group are then plotted. The intersection of the two histograms sets the passing score for that station. This is schematically represented in Figure 8. (Downing et al., 2006; J. J. Norcini, 2003)

For an examination that is made up of several stations, the passing score for the overall examination is calculated as the sum of the combined total stations’ passing scores.

The contrasting groups method has the advantage of being able to adjust the passing score to optimize the specificity and sensitivity of the passing score. This allows for the score to be set to minimize the risk of a false negative result or a false positive result.
2.9.8 Compensatory versus Conjunctive Model

A second issue with standard setting is deciding if the final pass/fail decision should be based on the overall score across all stations (compensatory model) or on passing a defined number of stations (non-compensatory model). (Newble, 2004) A compensatory model allows for an overall passing grade, with high examinee scores on some stations compensating for low scores on other stations. (Smee & Blackmore, 2001) For the purpose of a high stakes exam, a non-compensatory model has been suggested. For the Objective Structured Clinical Examination (OSCE) of the Licentiate of the Medical Council of Canada, a combination of both models is used, requiring that candidates pass both in terms of total score and minimum number of stations. (Dauphinee et al., 1997) This approach has been suggested for a high stakes exam.

**Figure 8.** Schematic representation of the contrasting groups standard setting methodology (Image adapted from de Montbrun S., Satterthwaite L. & Grantcharov T. 2015)
2.9.9 Reliability of the Passing Score

For credentialing or other high stakes examinations where a distinction between competent and non-competent is being made, the reliability of the pass/fail decision is of great importance. For theses mastery tests, the agreement co-efficient (consistently classifying examinees as competent or non-competent) is more important than the traditional test score reliability such as Cronbach’s alpha co-efficient \((r)\). This reliability is referred to as the agreement coefficient \((p_0)\) and as described by Subkoviak:

“...represents the proportion of examinees consistently classified on two administrations of a mastery test’”(Subkoviak, 1988)

This refers to the likelihood that an examinee would have the same pass/fail status if he or she took the exam a second time, assuming no gain in knowledge between administrations. Because administering an examination for a second time to be able to calculate this co-efficient is time consuming and typically not feasible, Subkoviak has outlined a methodology that can be used to estimate this co-efficient using data obtained from the administration of a single test.(Subkoviak, 1988)

Subkoviak has also recommended that the agreement co-efficient be at or above 0.85 for examinations that are being used to make serious, high stakes decisions.(Subkoviak, 1988)

2.9.10 Selecting a Standard Setting Methodology

There is no “gold standard” methodology that can be applied to establish a passing score. Many factors need to be considered when selecting a standard setting methodology.

Often the choice will depend on the structure of the examination, the feasibility of the methodology, the available expertise and the applicability of the standard setting methodology to
the type of data collected. For example, a performance based assessment that is based on the evaluation of observed examinee performance, participant-centered approaches make intuitive sense. These approaches allow the judgments to take place at the same time of the exam administration, thus making them efficient. Also, expert examiners tend to be familiar and comfortable with judging candidates’ performances. (Downing et al., 2006)

A major consideration is the intent of the outcomes of the assessment. For example, in an assessment establishing competence in the medical field, such as medical certification or graduation, a criterion based approach rather than a normative approach is most appropriate. (J. Norcini & Guille, 2002; J. J. Norcini, 2003)

It is recognized that different methodologies will produce different standards. However, in order for the standard to be credible, it must be set using appropriate judges, use a systematic method to set the score, demonstrate due diligence and produce reasonable outcomes. (J. Norcini & Guille, 2002), (J. Norcini & Shea, 1997)

One way to assess the reasonableness of the outcomes of the set passing scores is to explore the outcomes in terms of pass rates. For example, one would expect that the pass rate for a credentialing examination would be high for individuals who have completed a rigorous training program. (J. Norcini & Shea, 1997) Standard setting methodologies resulting in an unusually high failure rate would require investigation into the use of the standard setting methodology and its application and appropriateness to the data.

2.10 Summary

As a self-regulated profession, it is our duty to protect the public from harm by ensuring and formally documenting that graduating surgeons are technically competent individuals. While it is felt that the current North American training system has for decades produced a competent surgical workforce, formal documentation of the acquisition of technical skill through validated
examinations at the time of certification is currently lacking. The current surgical training milieu is highly focused on the achievement of domains of competence. The assessment of technical skill and establishing technical competence of trainees both during and at the time of completing residency is on the agenda for surgical training programs and national accreditation bodies.

This dissertation addresses several gaps in high stakes assessments of technical skill. Presently, high stakes technical skill examinations have not been developed and implemented into the surgical certification process. The PhD studies within this thesis apply contemporary validity frameworks to the development of high stakes technical skill exams at various key stages of development; from the development of an exam blueprint, to the implementation of the exam into board certification and finally the use of standard setting methodologies to set an exam pass score.
3 Chapter 3: Developing the blueprint for a general surgery technical skills certification exam: a validation study.

3.1 Abstract

**Background:** There is a recognized need to develop high stakes technical skills assessments for decisions of certification and resident promotion. High stakes exams requires a rigorous approach in accruing validity evidence throughout the developmental process. One of the first steps in development is the creation of a blueprint which outlines the potential content of exam. The purpose of this validation study was to develop an exam blueprint for a Canadian General Surgery Objective Structured Assessment of Technical Skill (GOSATS) certifying exam.

**Methods:** A Delphi methodology was used to gain consensus amongst Canadian general surgery program directors as to the content (tasks or procedures) that could be included in a certifying Canadian general surgery exam. Consensus was defined a priori as a Cronbach’s α ≥ 0.70. All procedures or tasks reaching a positive consensus (defined as ≥ 80% of program directors rated items as ≥ 4 on the 5-point Likert scale) were then included in the final GOSATS exam blueprint.

**Results:** Two Delphi rounds were needed to reach consensus. Of the 17 general surgery program directors across the country, 14 and 10 program directors responded to the first and second round respectively. A total of 59 items and procedures reached positive consensus and were included in the final GOSATS blueprint.

**Conclusions:** The present study has outlined the development of an exam blueprint for the GOSATS using a consensus based methodology. This validation study will serve as the foundational work from which simulated model will be developed, pilot tested and evaluated.
3.2 Introduction

The main goal of certification is to ensure that a candidate is competent in all facets that are required of their profession (Dcosta, 1986) and in doing so certify that the individual is safe to enter independent practice. The process of certification for surgeons within North America, however, has not included a formal or direct assessment of technical skill, which is clearly an essential domain of competence required of surgeons. Competency based training and assessment have become a major focus of surgical training around the globe. (Sonnadara et al., 2014; Szasz et al., 2014; The Royal College of Physician and Surgeons of Canada, 2014) However, high stakes board certification currently focuses on the assessment of knowledge and judgment (The Royal College of Physicians and Surgeons of Canada, 2014d, 2014e) while technical skill is evaluated typically with in-training evaluation reports which can be notoriously unreliable. (Feldman et al., 2004) Recognizing this gap in assessment at the time of certification, surgical boards have acknowledged the need to more formally document technical competence, and initiatives are underway to develop high stakes technical skills assessments for the purpose of certification. (S. L. de Montbrun et al., 2013)

The stakes of a credentialing examination are high, and the consequences of the results are significant. False positives occur when a non-competent candidate passes the examination, potentially putting patients at risk through lack of physician skill. False negatives occur when a competent candidate fails the examination, negatively impacting the candidate who has invested great efforts in training. (Downing, 2003) Because of the significance of the results of a certification exam, the validity evidence required for the interpretation of the test scores needs to be extensive, from multiple sources and collected on an on-going basis. (Korndorffer et al., 2010)

Validation studies, aim to build evidence for the argument that links performance on the assessment with the possession of the underlying construct being evaluated. A construct is an intangible entity that can be described but may not be easily measured. (Andreatta & Gruppen, 2009) For example, professionalism, knowledge and technical skill all represent constructs
within the domain of surgery. If a candidate does well on an examination, they likely possess more of the construct being tested than someone who performs poorly.

One of the first important steps in developing a high stakes assessment is to develop an exam blueprint, which outlines the content domain that will be covered and reflects the knowledge and skill expected of the profession. (Castle, 2002) Messick has described a contemporary validity framework, which is made up of five sources of validity evidence. The process of developing an exam blueprint, addressed one of these five sources – content validity – which aims to ensure that the content of the test parallels the domain being assessed. (Messick, 1989) Evidence of content validity can be accrued by demonstrating a structured approach to the development of the exam content; ensuring that the content mirrors the construct intended to be measured. An exam blueprint is an essential component and foundational necessity of a certification examination.

While platforms to assess technical skill have been developed, few have used a contemporary framework of validity in their development, few have used a rigorous methodology in the construction of exam content, and only one in North America has been developed for the purpose of certification. (S. L. de Montbrun et al., 2013; Ghaderi et al., 2015) The present study is a validation study, outlining one of the first key steps in developing a high stakes certification examination.

The specific aim of this validation study is to develop an examination blueprint for a certification technical skills examination for graduating Canadian general surgery residents (the General Surgery Objective Structured Assessment of Technical Skill – GOSATS) using a Delphi consensus methodology.

3.3 Methods

This study was approved by the research ethics board at the University of Toronto.
3.3.1 Delphi Methodology

The Delphi technique was selected to attain consensus on the content of a technical skills exam to be developed for Canadian graduating general surgery resident for the purpose of certification.

The Delphi is an iterative, multistage process that aims to achieve consensus amongst a group of experts through a series of structured questionnaires.(Hasson, Keeney, & McKenna, 2000) The four major cornerstones of the Delphi techniques adhered to in this study include the use of and expert panel, subject anonymity, multiple rounds, and statistical aggregation of results.(Hasson et al., 2000; Hsu, 2007; RAND, 1976)

3.3.1.1 Expert Panel

In order to gain consensus on a specialty topic, it has been recommended that the expert panel for a Delphi include 10-20 individuals who are felt to be highly trained and competent in the specific domain being studied. (Delbecq, Van de Ven, & Gustafson, 1975) Content experts in the field of general surgery training participated in this study. With their unique expertise in general surgery training, Canadian general surgery program directors from the 17 Royal College of Physicians and Surgeons of Canada (RCPSC) accredited programs were invited to participate.

3.3.1.2 Anonymity

Anonymity limits the influence that dominant individuals or dominant views may have on other participants’ responses.(Goodman, 1987) This was achieved in our study by the use of an online platform, which eliminated any face to face contact amongst the expert panel.
3.3.1.3 Delphi Rounds

The Delphi technique requires a minimum of two rounds; the first round being open ended and subsequent rounds being closed ended. (Jairath & Weinstein, 1994; Powell, 2002) As suggested in the literature, the first iteration of the present Delphi consisted of open ended questions. During this round, expert participants were asked to list any general surgery task or procedure that could be included in a final assessment of technical skill for graduating Canadian general surgery residents grouped into 8 anatomic categories (open and minimally invasive upper gastrointestinal tract, open and minimally invasive lower gastrointestinal tract, open and minimally invasive hepatopancreatobiliary, open and minimally invasive hernias, perianal, breast, emergency and trauma, soft tissue). The second iteration consisted of this list of tasks and procedures generated by the expert panel supplemented by tasks and procedures added through a review of the objectives of general surgery training outlined by the RCPSC. (Objectives of Training in the specialty of General Surgery, 2010)

The second Delphi round, which was closed ended, listed each item within the 8 categories all of which were ranked on a 5-point Likert scale. The scale avoided a neutral ranking point with the values anchored as follows: 1) unimportant 2) less important 3) somewhat important 4) important 5) very important. Responses were confidential.

3.3.1.4 Aggregation of Results and Definition of Consensus

The data from responses to the second closed ended Delphi round were then statistically analyzed to generate a median and interquartile range (IQR) for each identified item within each anatomic category.

The use of Cronbach’s $\alpha$ as a measure of internal consistency for the Delphi consensus technique in the domain of medicine has been previously established. (Graham et al., 2003) A Cronbach’s $\alpha$ of $\geq 0.70$ has been recommended for consensus for research purposes (Bland & Altman, 1997) and was selected a priori as the level of consistency required for this study.
3.3.2 Inclusion Criteria for GOSATS Blueprint

Once consensus had been reached for each anatomic category, items from each category were then selected for inclusion in the final GOSATS exam blueprint. Items were deemed to have reached positive agreement and thus selected for inclusion in the blueprint if ≥ 80% of respondents ranked items as either a 4 (important) or 5 (very important). Negative agreement was defined as >80% of responses ranked as either 1 (unimportant) or 2 (less important). Neutral agreement was defined as all other responses. Both negative and neutral items were excluded from the GOSATS blueprint.

3.3.3 Administration

Prior to the administration of each Delphi round, the questions were piloted within our research group to ensure clarity and ease of use. The online platform SurveyMonkey™ was used to administer the Delphi. Program directors were sent a link to the website through their publicly available University email addresses. A total of two administrations were required to achieve consensus. Each Delphi round closed after 6 weeks and 3 email reminders. The second round was only sent to those that had responded to the first round.

3.4 Results

3.4.1 Participants

Of the 17 general surgery program directors across the country, 14 and 10 program directors responded to the first Delphi round and second Delphi round respectively. Geographic representation was broad and included programs from across the country (Table 5). Furthermore, the program directors demonstrated a range of sub specialty training (Table 5).
3.4.2 Consensus

Consensus was achieved after two Delphi rounds. After two rounds the Cronbach’s $\alpha$ for the anatomic categories ranged from 0.72-0.92. Six of the eight anatomic categories reached a Cronbach’s $\alpha \geq 0.80$.

3.4.3 Final GOSATS Exam Blueprint

A total of 59 items and procedures reached positive consensus and were included in the final GOSATS exam blueprint. All procedures and tasks included in the final GOSATS exam blueprint are outlined in Table 6 along with their medians and interquartile ranges.
Table 5. Participating general surgery programs and subspecialty of program director

<table>
<thead>
<tr>
<th>General Surgery Program</th>
<th>Subspecialty</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of British Columbia</td>
<td>surgical oncology</td>
</tr>
<tr>
<td>University of Alberta</td>
<td>trauma/acute care surgery</td>
</tr>
<tr>
<td>University of Calgary</td>
<td>surgical oncology</td>
</tr>
<tr>
<td>University of Saskatchewan</td>
<td>critical care medicine</td>
</tr>
<tr>
<td>Western University</td>
<td>colorectal surgery &amp; trauma/acute care surgery</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>trauma/acute care surgery &amp; critical care medicine</td>
</tr>
<tr>
<td>Memorial University</td>
<td>minimally invasive surgery</td>
</tr>
<tr>
<td>University of Ottawa</td>
<td>colorectal surgery</td>
</tr>
<tr>
<td>Sherbrooke University</td>
<td>minimally invasive surgery</td>
</tr>
<tr>
<td>Dalhousie University</td>
<td>hepatopancreatobiliary surgery &amp; transplant</td>
</tr>
<tr>
<td>McMaster University*</td>
<td>hepatopancreatobiliary surgery</td>
</tr>
<tr>
<td>University of Montreal*</td>
<td>surgical oncology</td>
</tr>
<tr>
<td>Queens University*</td>
<td>hepatopancreatobiliary surgery</td>
</tr>
<tr>
<td>University of Manitoba*</td>
<td>trauma/acute care surgery</td>
</tr>
</tbody>
</table>

*program directors that responded to the first Delphi round only
Table 6. Positive consensus on procedures that could be included in a general surgery certification examination

<table>
<thead>
<tr>
<th>Anatomical categorization of tasks and procedures</th>
<th>Cronbach’s $\alpha$ per question</th>
<th>Median (IQR*) per item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Open and minimally invasive upper gastrointestinal tract</strong></td>
<td>0.81</td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Open distal gastrectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oversewing of bleeding duodenal ulcer</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Graham patch for perforated duodenal ulcer</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>OGD</td>
<td>5.0 (1.0)</td>
<td></td>
</tr>
<tr>
<td>Exposure of pancreas</td>
<td>4.0 (1.0)</td>
<td></td>
</tr>
<tr>
<td><strong>2. Open and minimally invasive lower gastrointestinal tract</strong></td>
<td>0.90</td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>Open appendectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laparoscopic appendectomy- simple</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Laparoscopic appendectomy- complicated</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Sigmoidoscopy</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Colonoscopy</td>
<td>5.0 (1.0)</td>
<td></td>
</tr>
<tr>
<td>Laparoscopic right hemicolecction</td>
<td>4.0 (1.0)</td>
<td></td>
</tr>
<tr>
<td>Open right hemicolecction</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Open sigmoid resection</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Open upper rectal resection</td>
<td>4.0 (1.0)</td>
<td></td>
</tr>
<tr>
<td>Hartman procedure</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Colostomy</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Ileostomy</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td><strong>3. Open and minimally invasive hepatopancreatobiliary</strong></td>
<td>0.82</td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>Open cholecystectomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Laparoscopic cholecystectomy – simple</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Laparoscopic cholecystectomy – complicated</td>
<td>5.0 (0)</td>
<td></td>
</tr>
<tr>
<td>CBD exploration</td>
<td>4.0 (0)</td>
<td></td>
</tr>
<tr>
<td>Kocherize duodenum</td>
<td>5.0 (0)</td>
<td></td>
</tr>
</tbody>
</table>
Table 6 continued

<table>
<thead>
<tr>
<th>Anatomical categorization of tasks and procedures</th>
<th>Cronbach’s α per question</th>
<th>Median (IQR) per item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4. Open and minimally invasive hernias</strong></td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Open mesh repair of inguinal hernia</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Open tissue repair of inguinal hernia</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Recurrent inguinal hernia</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Femoral hernia repair</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Open umbilical hernia repair</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Open incisional hernia</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Laparoscopic incisional hernia</td>
<td></td>
<td>4.0 (0)</td>
</tr>
<tr>
<td>Parastomal hernia</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td><strong>5. Perianal</strong></td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Hemorrhoidectomy</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Hemorrhoid banding</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Lateral internal sphincterotomy</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Drainage of perianal sepsis</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Simple fistulotomy</td>
<td></td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>Placement of seton drain</td>
<td></td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>EUA</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td><strong>6. Breast</strong></td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Duct excision</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Breast lumpectomy</td>
<td></td>
<td>5.0 (0.25)</td>
</tr>
<tr>
<td>Simple mastectomy</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Modified radical mastectomy</td>
<td></td>
<td>5.0 (0)</td>
</tr>
<tr>
<td>Skin sparing mastectomy</td>
<td></td>
<td>4.0 (1.0)</td>
</tr>
<tr>
<td>Axillary dissection</td>
<td></td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>Sentinel lymph node biopsy</td>
<td></td>
<td>5.0 (1.0)</td>
</tr>
<tr>
<td>Needle loc biopsy</td>
<td></td>
<td>5.0 (0)</td>
</tr>
</tbody>
</table>
Table 6 continued

<table>
<thead>
<tr>
<th>Anatomical categorization of tasks and procedures</th>
<th>Cronbach’s α per question</th>
<th>Median (IQR) per item</th>
</tr>
</thead>
</table>

7. Emergency and trauma  
0.90

- Damage control laparotomy  
- Trauma splenectomy  
- Control of liver trauma hemorrhage  
- Suture repair of bowel injury  
- Surgical management of SBO  
- Surgical management of LBO  
- Surgical management of ischemic bowel  
- Emergency surgical airway  
- Chest tube  
- Central line

8. Soft tissue  
0.74

- Debridement of necrotizing soft tissue infection  
- Resection of benign skin and soft tissue lesion  
- Resection of malignant skin lesion  
- Lymph node biopsy

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

OGD – Esophagogastroduodenoscopy, CBD – common bile duct, EUA - examination under anesthesia, LBO – large bowel obstruction, SBO – small bowel obstruction
3.5 Discussion

Developing a high stakes examination such as certification requires a structured, rigorous process guided by contemporary validity theory. The present study used a Delphi methodology to gain consensus amongst Canadian general surgery program directors on the potential content of a technical skills examination for graduating Canadian general surgery resident, using a contemporary validity theory. Consensus amongst program directors was achieved and a blueprint consisting of 59 potential tasks and procedures was developed for the GOSATS exam. This first foundational step is important in exam development.

Establishing the achievement of competence is on the agenda for both surgical training programs (University of Toronto Faculty of Medicine, 2012), national accreditation bodies and surgical societies.(Canada, 2013; S. L. de Montbrun et al., 2013) Currently technical competence of general surgery residents around the world is not formally assessed at the time of certification. However, it is becoming increasingly apparent that assessing technical skill is important with evidence suggesting that poor technical skill can result in increased postoperative complications, readmission rates and emergency department visits.(Birkmeyer et al., 2013)

A tool to assess technical competence in a high stakes setting such as certification requires a high level of validity evidence, because the consequences of this type of exam are significant. Validation studies are important throughout the development process of a high stakes examination, such as certification, in order to build an argument for the valid interpretation of test scores.

One of the first steps in developing a certification examination is the development of an exam blueprint. This blueprint is intended to outline the content domain that should be included in the assessment and reflects the expected knowledge and skill of the profession.(Castle, 2002)
rigorous approach to the development of the exam blueprint provides a solid foundation for the exam, and addresses one aspect of validity within Messick’s framework.

Messick provides a robust framework to understand validity evidence. This framework can be used to develop validation studies; helping to guide study design to gather validity evidence from the five various sources of validity evidence. The ‘content’ domain of Messick’s theory addresses the issue of ensuring that there is a relationship between the content of the exam and the construct being measured. (Messick, 1989) A validation study to develop an exam blueprint for an assessment tool is one way to build content validity evidence. (Ghaderi et al., 2015) The present study used Messick’s framework, to build initial evidence of content validity of the GOSATS through the development of an examination blueprint using a Delphi consensus methodology.

The only North American surgical body to move forward with developing this type of high stakes technical skills assessment has been the American Society of Colon and Rectal Surgeons (ASCRS), who have developed the COSATS exam. (S. L. de Montbrun et al., 2013) This exam has been developed for the purpose of certification, and content validity was sought by gaining expert consensus amongst members of the operative competency committee of the ASCRS as to the tasks and procedures that would be expected of a colorectal surgeon entering independent practice. This list was then supplemented with a review of the American Board of Colon and Rectal Surgery objectives of colorectal surgery training and the list of minimum procedures required during colorectal fellowship training. (Education, 2015b) In addition to content validity evidence, further validity evidence has been accrued through further validation studies. In 2014, the Colorectal Objective Assessment of Technical Skills (COSATS) exam was a mandatory component of board certification. The COSATS exam appears to be testing a different construct than the written or oral board examination, with the COSATS exam identifying deficiencies in technical skill in individuals who, with the current board certification process, would go on to be certified; (S. de Montbrun, Roberts, Satterthwaite, & MacRae, 2016) highlighting the importance of testing technical skill at the time of board certification. The development and study of the COSATS has paved the way for other surgical specialties to move forward with developing simulated technical skills examinations for high stakes assessment.
Others have determined the operative procedures required of general surgery graduates. Bell et al. surveyed US general surgery program directors to establish which general surgical procedure are considered to be essential to general surgical practice.(Bell et al., 2009) Our study adds to this literature in several ways. Firstly, the purpose from the outset of the present study was to develop a blueprint of tasks for a high stakes examination. Secondly, the present study was aimed at achieving consensus on procedures to be included in a certification examination rather than classifying procedures according to the level of competence required for each. Lastly, because the exam is currently being developed for Canadian general surgery training, it focused exclusively on Canadian training with the use of Canadian general surgery program directors as experts.

Developing an examination for general surgery would have far reaching implications given the size of the target audience. The annual general surgery cohort graduating in both Canada and the United States is significantly larger than the annual cohort of colorectal graduates. For example, within Canada there are approximately 85 general surgery graduates a year,(service, 2016) and only approximately 7 colorectal graduates a year. Within the United States, for the academic year of 2015-2016, as reported by the Accreditation Council for Graduate Medical Education (ACGME), there were 262 accredited general surgery programs in the United States, with a total of 7907 residents on duty.(Accreditation Council for Graduate Medical Education, 2016) Assuming a 5 year general surgery program, this would be on average a total of 1581 residents per year. This is in contrast to the approximate 100 colorectal graduates a year in the United States. An assessment tool to assess general surgery graduates could have profound near-term impact, helping to identify those residents who may not be prepared for independent practice.

This study is the first step in developing a general surgery objective structured assessment of technical skill. Simulated models are currently in the process of being developed and a pilot administration of the GOSATS exam is planned for the near future.
3.6 Conclusions

A rigorous approach is required when developing a high stakes examination; with ongoing validation studies building evidence for interpretation of test scores. The present study outlines the development of an exam blueprint for a general surgery technical skills certification exam. The development of a blueprint is an essential first step in developing a high stakes examination in accordance with Messick’s framework of validity. This validation study used a Delphi consensus methodology to outline content for the GOSATS examination. This blueprint will guide the development of simulated models for the GOSATS examination and serve as an exam bank for future GOSATS stations.
Chapter 4: Setting passing scores for technical performance in surgery: lessons learned from 10 years of resident assessment

This chapter has been modified from the following:


4.1 Abstract

**Background:** Competency based assessment is a paradigm shift in surgical training. One of the major challenges is defining a passing score for competent technical performance. The objective of this study was to set passing scores for the Objective Structured Assessment of Technical Skill (OSATS).

**Methods:** A retrospective analysis of prospectively collected OSATS performance data from 513 PGY 1 surgical residents from 2002-2012 was used for this study. Three standard setting methodologies (contrasting groups, borderline group, borderline regression method) were applied to set passing scores for each OSATS technical task. A compensatory and conjunctive model were used to assign an overall exam pass/fail status to general surgery residents. The consistency of the pass/fail status was compared across the three methodologies and between a compensatory and a conjunctive model.

**Results:** The passing scores for each OSATS technical task were stable across all three methodologies, with very little variation. The consistency of the pass/fail decision across the three methods was 95.5% and 93.2% using a compensatory and conjunctive model respectively. The consistency across a compensatory and conjunctive model was also very high with only 2.25% of individuals having an inconsistent status.
Conclusion: This is the first large scale study to introduce and apply standard setting methodologies using the world’s largest OSATS database and the first to demonstrate both the feasibility and consistency of the three methodologies. These results can help guide exam administrators in selecting and implementing the most feasible and appropriate standard setting methodology for their assessment tool.
4.2 Introduction

Assessing surgical skill and developing objective ways to evaluate competence has become a focus of surgical education. (Hampton, 2015) With this trend, competency based training has become mandated by regulatory bodies and is being incorporated into residency programs. (Orthopedic curriculum and assessment project, 2006; Ritchie, 2001; The Royal College of Physician and Surgeons of Canada, 2013a) Achieving competency standards aims to demonstrate that trainees and surgeons have the necessary skills to provide safe patient care. Competency standards have been established in the domains of knowledge and judgment, however the evidence on standards of performance for technical skills assessment is still insufficient. (Fraser et al., 2003; Szasz et al., 2014)

Over the past several years, validated, reliable and objective tools to assess technical skill have been developed and implemented in practice, however, setting passing scores to establish the achievement of competency is limited. (J. Beard et al., 2009; S. L. de Montbrun et al., 2013; Martin et al., 1997; Szasz et al., 2014; Vassiliou et al., 2006) The definition of competence in the literature typically reflects an individual who demonstrates a minimum standard of performance safe for independent practice; (Szasz et al., 2014) however, this conceptualization has rarely been translated into a passing score, thus limiting the ability of these assessments to differentiate between a competent and non-competent trainee. A passing score is a fundamental component of competency based programs, acting as a cut point along the score scale that defines the boundary between someone who performs ‘well enough’ and someone who does not. (Cusimano, 1996; J. J. Norcini, 2003) Even major competency based education programs that have been incorporated into training have not rigorously addressed this key component of assessment. (Ferguson et al., 2013; Sonnadara et al., 2014)

As competency based programs continue to shift the surgical education paradigm, insight into establishing passing scores for technical performance in surgery is essential: selecting, implementing and comparing these methodologies in order to produce feasible, credible and defensible passing scores is imperative. (Schindler et al., 2007)
One of the best validated technical skills assessments is the Objective Structured Assessment of Technical Skill (OSATS), developed at the University of Toronto as a platform to assess technical skills in a simulated setting. (Martin et al., 1997) Studies exploring the validity and reliability of this exam have shown that it has the potential for use in high stakes assessments such as certification or resident promotion; however, one of its main limitations is its lack of an established passing score. (R. K. Reznick & MacRae, 2006)

The University of Toronto has administered the OSATS exam to all first year surgical residents for over the past decade and maintains the world’s largest OSATS performance database. This database overcomes the common limitation of small numbers found in the surgical assessment literature and provides ample data for the application and study of several standard setting methodologies.

Recognizing the need to study standard setting in the context of surgical performance assessment, the purpose of this study was (1) to set a passing score for each OSATS technical skill task, using three standard setting methodologies, (2) to evaluate the best way to establish an overall pass/ fail criteria for the OSATS exam as a whole, and (3) to assign an overall pass/ fail status to examinees and compare this status across the three methodologies.

4.3 Methods

The research ethics board at St. Michael’s Hospital approved this study. At the University of Toronto, surgical residents from all specialties take the OSATS exam during their postgraduate year (PGY) 1. The OSATS exam consists of 8 simulated technical skills tasks performed in the skills lab. Examinees rotate through all 8 tasks and are evaluated by an expert examiner using a task specific checklist, global rating scale (GRS), and overall performance scale. (Martin et al., 1997) The task specific checklist itemizes each step of the task and was assessed using a binary scale: done correctly or not done/incorrectly. The GRS used was the original OSATS 7 category Likert scale allowing for a maximum score of 35. The overall performance scale rated trainees from 1-5 (1 being very poor and 5 being clearly superior) Figure 9.
Expert examiners consisted of faculty surgeons from all surgical divisions. The examiners received instructions on examining several weeks prior to the OSATS, then underwent a 30 minute orientation and training session on the day of the exam.

OSATS data was prospectively collected and retrospectively reviewed from all PGY1 surgical residents from 2002 to 2012 at the University of Toronto (n= 513). Over the ten-year study period, 18 different technical skills stations have been administered. Global rating scales have shown to be more reliable and appropriate than checklists for summative assessments; (Regehr et al., 1998) therefore, only the GRS data as well as the overall performance score was used to set the passing scores for each technical task.

4.3.1 Establishing Passing Scores for Each OSATS Task

Three standard setting methodologies (contrasting groups method (CG), the borderline group method (BG), borderline regression method (BRM)) were used to set a passing score for each OSATS task. The passing scores generated by the three methodologies were compared for consistency in order to establish initial evidence of credibility of the scores. (Cizek & Bunch, 2007)
GLOBAL RATING SCALE OF OPERATIVE PERFORMANCE

Please circle the number corresponding to the candidate’s performance regardless of the candidate’s level of training.

<table>
<thead>
<tr>
<th>Category</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respect for tissue</td>
<td>Frequently used unnecessary</td>
<td>Careful handling of tissue but</td>
<td>Consistently handled tissue</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>force on tissue or caused damage</td>
<td>occasionally caused inadvertent</td>
<td>appropriately with minimal</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>by inappropriate use of</td>
<td>damage</td>
<td>damage to tissue</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>instruments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time and motion</td>
<td>Many unnecessary moves</td>
<td>Efficient time/motion but some</td>
<td>Clear economy of movement</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>unnecessary moves</td>
<td>and maximum efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instrument Handling</td>
<td>Repeatedly makes tentative or</td>
<td>Competent use of instruments</td>
<td>Fluid movements with instruments</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>awkward moves with instruments</td>
<td>but occasionally appeared</td>
<td>and no stiffness or awkwardness</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>through inappropriate use</td>
<td>stiff or awkward</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of Instruments</td>
<td>Frequently asked for</td>
<td>Knew names of most instruments</td>
<td>Obviously familiar with</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>wrong instrument or used</td>
<td>and used appropriate instrument</td>
<td>instruments and their names</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>inappropriate instrument</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow of Operation</td>
<td>Frequently stopped</td>
<td>Demonstrated some forward</td>
<td>Obviously planned course of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>operating and seemed unsure of</td>
<td>planning with reasonable</td>
<td>operation with effortless flow</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>next move</td>
<td>progression of procedure</td>
<td>from one move to the next</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of Assistants</td>
<td>Consistently placed assistants</td>
<td>Appropriate use of assistants</td>
<td>Strategically used assistants to</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>poorly or failed to use</td>
<td>most of the time</td>
<td>the best advantage at all times</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>assistants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of Specific Procedure</td>
<td>Deficient knowledge.</td>
<td>Knew all important steps of</td>
<td>Demonstrated familiarity with</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Required specific instruction</td>
<td>operation</td>
<td>all steps of the operation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at most steps of operation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OVERALL PERFORMANCE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very poor</td>
<td>Competent</td>
<td>Clearly superior</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 9. OSATS Global rating scale
4.3.1.1 The Contrasting Groups Methodology

The contrasting groups method divides the cohort into two groups: those who clearly ‘pass’ and those who clearly ‘fail’ the exam. (Livingston & Ziesky, 1982) Histograms of scores from the ‘pass’ group and the ‘fail’ group are then plotted and the passing score is set at the intersection of these two distributions (Figure 10). (Downing et al., 2006; J. J. Norcini, 2003)

For the analysis, individuals who scored a 1 or 2 on the overall performance scale were categorized as the ‘fail’ group, and individuals who scored a 4 or 5 were categorized as the ‘pass’ group. The intersection of these two plotted histograms set the passing score for the task.

![Passing score diagram](image)

**Figure 10.** The contrasting groups standards setting methodology.
4.3.1.2 Borderline Group Methodology

The borderline group method sets the passing score at the level of the ‘borderline candidate’, (Livingston & Ziesky, 1982) who is conceptualized as an individual sitting right on the ‘border’ of the passing score. (Boulet, De Champlain, & McKinley, 2003; Sturmburg & Hinchy, 2010)

For this study, all candidates who scored a 3 on the overall performance scale (described as ‘competent’) made up the ‘borderline’ group. The passing score for each technical skills task was the calculated mean GRS score for this ‘borderline’ group (Figure 11).

4.3.1.3 Borderline Regression Methodology

The borderline regression method also centers on the idea that the passing score is set at the level of a ‘borderline candidate’. However, rather than using a subset of data points, this method allows for data from all examinees to be used in the analysis by implementing a linear regression model.

To establish the pass score for each task, GRS scores from all examinees were regressed onto their overall performance score (Figure 12). The point along the overall performance scale defining a ‘competent’ performance (score of 3) was inserted into the linear equation to establish the passing score.
Figure 11. The borderline group standard setting methodology

Figure 12. Borderline regression standard setting methodology
4.3.2 Establishing a Pass Criteria for the Overall Exam

In addition to setting passing scores for each OSATS task, two different methods were used to establish an individuals’ pass/ fail status for the overall exam (all 8 stations). Passing criteria can be established for the exam using a compensatory or a conjunctive model. A compensatory model allows an examinee to pass the OSATS exam if they achieve the passing score, even if they fail some tasks (allowing high scores on some tasks to compensate for low scores on other tasks). (Newble, 2004) In contrast, a conjunctive model requires that candidates not only achieve an overall passing score but also pass a certain number of tasks or certain required tasks. In high stakes exams such as certification or promotion, a conjunctive model has been recommended in order to ensure that students demonstrate competence in crucial or essential tasks or domains. (Dauphinee et al., 1997) In this study a compensatory and conjunctive model were compared in order to evaluate consistency across the two models; with a conjunctive model candidates were required to pass based on overall score in addition to passing at least 50% of the tasks.

Both a compensatory and conjunctive models were applied to all three standard setting methodologies. Due to the variation in the 8 technical tasks used each year, the final OSATS exam passing score for each methodology was calculated as the sum of the combined 8 tasks’ passing scores. This analysis was done for general surgery residents only (n=133) because of an ongoing study exploring the predictive ability of the passing scores using this cohort.

4.4 Results

4.4.1 Establishing Passing Scores for Each OSATS Task

The 18 technical skills tasks, their passing scores, and the number of resident observations included in the analysis are outlined in Table 7.
<table>
<thead>
<tr>
<th>Technical Skills Task</th>
<th>Passing Score Contrasting Groups Method (SD fail group, SD pass)¹</th>
<th>Passing Score Borderline Group Method (SD)²</th>
<th>Passing Score Borderline Regression Method (95% CI)³</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal closure (n=222)</td>
<td>20.62 (2.26, 2.44)</td>
<td>19.80 (1.972)</td>
<td>20.06 (19.05-21.07)</td>
</tr>
<tr>
<td>Forearm cast (n=39)</td>
<td>**</td>
<td>19.00 (1.706)</td>
<td>19.87 (16.83-22.90)</td>
</tr>
<tr>
<td>Bowel anastomosis 1 (n=248)</td>
<td>19.32 (2.29, 2.98)</td>
<td>19.01 (1.958)</td>
<td>19.21 (18.39-20.03)</td>
</tr>
<tr>
<td>Bowel anastomosis 2 (n=295)</td>
<td>19.99 (2.25, 3.29)</td>
<td>20.03 (2.820)</td>
<td>20.00 (18.89-21.10)</td>
</tr>
<tr>
<td>Chest tube (n=504)</td>
<td>19.97 (2.52, 2.77)</td>
<td>19.18 (2.186)</td>
<td>19.48 (18.86-20.09)</td>
</tr>
<tr>
<td>Control IVC hemorrhage (n=96)</td>
<td>21.28 (2.92, 3.55)</td>
<td>21.50 (3.273)</td>
<td>21.58 (19.82-23.33)</td>
</tr>
<tr>
<td>Bronchoscopy (n=30)</td>
<td>**</td>
<td>19.00 (1.927)</td>
<td>18.83 (15.47-22.19)</td>
</tr>
<tr>
<td>Foley (n=92)</td>
<td>23.31 (3.21, 2.43)</td>
<td>22.56 (2.068)</td>
<td>22.30 (20.82-23.77)</td>
</tr>
<tr>
<td>Insertion T–tube (n=43)</td>
<td>21.63 (2.77, 4.07)</td>
<td>20.00 (2.517)</td>
<td>20.99 (18.80-23.19)</td>
</tr>
<tr>
<td>Lap tying 2 (n=137)</td>
<td>18.43 (2.04, 2.85)</td>
<td>18.47 (2.311)</td>
<td>18.77 (17.69-19.86)</td>
</tr>
<tr>
<td>Lap tying 1 (n=96)</td>
<td>19.13 (3.10, 2.76)</td>
<td>18.73 (1.818)</td>
<td>18.85 (17.89-19.82)</td>
</tr>
<tr>
<td>Lap vessel dissection (n=86)</td>
<td>18.39 (4.05, 3.45)</td>
<td>17.56 (2.464)</td>
<td>18.17 (16.64-19.70)</td>
</tr>
<tr>
<td>ORIF (n=413)</td>
<td>19.02 (3.21, 3.13)</td>
<td>18.48 (1.970)</td>
<td>18.69 (18.14-19.23)</td>
</tr>
<tr>
<td>Skin lesion excision (n=421)</td>
<td>20.74 (2.62, 2.96)</td>
<td>19.11 (2.320)</td>
<td>19.68 (18.94-20.43)</td>
</tr>
<tr>
<td>STSG (n= 116)</td>
<td>18.51 (2.67, 3.08)</td>
<td>19.09 (2.156)</td>
<td>18.82 (17.79-19.84)</td>
</tr>
<tr>
<td>Tracheostomy (n=513)</td>
<td>20.03 (2.81, 3.61)</td>
<td>19.48 (2.533)</td>
<td>19.77 (18.98-20.56)</td>
</tr>
<tr>
<td>Vascular anastomosis (n=295)</td>
<td>21.69 (2.91, 2.40)</td>
<td>21.09 (2.116)</td>
<td>21.19 (20.28-22.10)</td>
</tr>
<tr>
<td>Z plasty (n=435)</td>
<td>20.32 (2.87, 2.84)</td>
<td>19.31 (2.393)</td>
<td>19.57 (18.89-20.25)</td>
</tr>
</tbody>
</table>

¹ Standard deviation of the mean of the fail group and the standard deviation of the mean of the pass group

² Standard deviation of the mean of the borderline group

³ Unable to calculate due to too few candidates in the fail group

⁴ 95% confidence interval for the slope of the of the regression line

IVC – inferior vena cava, ORIF – open reduction internal fixation, STSG – split thickness skin graft
The range of GRS passing scores for the 18 technical tasks was 18.39–23.31 (52.29%–66.6%) for the contrasting groups, 17.56 – 22.56 (50.17%–64.46%) for the borderline group, and 18.1 – 22.3 (51.91%–63.71%) for the borderline regression method. The passing scores for each technical task were stable across all three standard setting methodologies, with very little variation (Figure 13). A passing score for two stations (bronchoscopy and forearm cast) could not be calculated for the contrasting groups methodology because of too few individuals in the ‘fail’ group.

4.4.2 Establishing a Pass Criteria for the Overall Exam

In the compensatory model (pass based on score alone), the consistency of the pass/ fail decision across the three methods was 95.5% with only 6/133 (4.5%) individuals having an inconsistent status (Table 8).

In the conjunctive model (pass based on score and passing ≥50% of the tasks), the consistency of the pass/fail decision across the three methods was also very high, at 93.2%, with only 9/133 (6.8%) individuals having an inconsistent status (Table 8).

When the pass/ fail decisions of the compensatory and conjunctive models were compared, the two models demonstrated very little difference, with only 3/133 (2.25%) having an inconsistent status.
Figure 13. Passing scores for all three standards setting methodologies
Table 8. Consistency of the pass/ fail status across three methodologies

<table>
<thead>
<tr>
<th></th>
<th>Compensatory Model</th>
<th></th>
<th>Conjunctive Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number (n)</td>
<td>Percent (%)</td>
<td>Number (n)</td>
<td>Percent (%)</td>
</tr>
<tr>
<td>Consistent pass‡</td>
<td>106</td>
<td>79.7</td>
<td>102</td>
<td>76.7</td>
</tr>
<tr>
<td>Consistent fail*</td>
<td>21</td>
<td>15.8</td>
<td>22</td>
<td>16.5</td>
</tr>
<tr>
<td>Inconsistent status</td>
<td>6</td>
<td>4.5</td>
<td>9</td>
<td>6.8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>133</strong></td>
<td><strong>100</strong></td>
<td><strong>133</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

‡ pass status for all three methodologies
* fail status for all three methodologies

4.5 Discussion

Competency based training and assessment requires not only the availability of valid and reliable tools but also the establishment of credible passing scores defining the boundary between a competent and non-competent performance. Competence in technical skill has been defined as being “suitable, fit or adequate”; (Satava et al., 2003) however, operationalizing this definition into a passing score in order to differentiate those who are competent from those who are not has been a major limitation in technical skills assessment. Tools to assess technical skill have been developed, validated and implemented in training programs, (Faurie & Khadra, 2012) however many fail to set a passing score, limiting their ability to make decisions of competency. While the process of setting standards has been well established in written and oral examinations, the application of standard setting methodologies in technical skills assessment has been under investigated and is acknowledged as an area deserving further study.

The present study addressed this gap in knowledge by applying and comparing three standard setting methodologies to a validated technical skills assessment, comparing both a compensatory and conjunctive model. Our results indicate that the contrasting groups, borderline group and
borderline regression methods are feasible and can be easily applied to OSATS data. More importantly, the consistency seen across the three methods builds initial evidence of validity and credibility of the set passing scores. (Cizek & Bunch, 2007) Furthermore, the consistency of the cut scores suggests that any of the three methodologies can be used to set passing scores for OSATS performance data, allowing the choice to be based on the capabilities of the assessment infrastructure, local expertise and exam size. This study also demonstrated little difference in outcome between a compensatory and conjunctive model, suggesting that either model could be implemented, taking into account the goal of the evaluation tool. The results of this study are timely, because setting defensible and credible passing scores is a key component of competency based education.

Only a few studies have set standards of performance for technical skill. The best example comes from the technical skills component of the Fundamentals of Laparoscopic Surgery (FLS) program. (Fraser et al., 2003) A contrasting groups method was used, with receiver operator curves identifying a score with the highest sensitivity and specificity. (Fraser et al., 2003) This passing score allows for judgments of competence and has allowed the FLS to become a component of the certification process for the American Board of Surgery exam. (Surgery, 2012) A passing score for FLS, however, was set using a single methodology, and a compensatory model. Using a single methodology limits the ability to compare results across several methods; where the consistency in passing scores across different methods builds evidence of validity of the standards. In addition to using a single methodology, most studies on technical skills standard setting have focused on single procedures or task, rather than a broad domain of technical skill (J. D. Beard, Education, Training Committee of the Vascular Society of Great, & Ireland, 2005; Jelovsek et al., 2010). In contrast, the present study uses a rigorous approach to setting performance standards, comparing three different methodologies, assessing a broad range of technical skill and also comparing a compensatory and conjunctive model to ensure the scores are defensible and credible.

While standard setting in technical skills assessment is limited, one of the best examples of the use and implementation of standard setting in performance based assessment comes from the objective structured clinical exam (OSCE) literature. (J. J. Norcini, 2003) OSCE examinations
have studied and compared different standard setting approaches (Burrows et al., 1999; Clauser & Clyman, 1994; Cusimano & Rothman, 2004; Dauphinee et al., 1997; Jalili et al., 2011; Kaufman et al., 2000; Kramer et al., 2003; Livingston & Zieky, 1989; Medical Council of Canada, 2013b; Morrison et al., 1996; Schoonheim-Klein et al., 2009; United States Medical Licensing Examination, 2014; Wilkinson et al., 2001; Wood et al., 2006) and demonstrated evidence of valid and credible passing scores. Having set defensible pass/fail standards has allowed the OSCE to be incorporated into high stakes decisions; for certification and licensure. (Medical Council of Canada, 2013b; United States Medical Licensing Examination, 2014) Though the OSCE exam does not assess technical skill, it does have significant structural similarities, allowing this body of literature to inform standard setting as it applies to the OSATS exam.

From the OSCE literature, several methodologies have been suggested as appropriate for performance based assessments in general. (Downing et al., 2006) Though there is no “gold standard” approach, it is suggested that a criterion based approach is most appropriate as it establishes a student’s level of competence based on a specified standard of achievement rather than on the performance of other members in the group (normative based). (Barman, 2008; Downing, 2003) Furthermore, criterion based standard setting is more appropriate in the setting of high stakes examinations such as medical or surgical examinations where decisions of competence and licensure are being made. (McIlhenny & Orr, 2002; J. J. Norcini, 2003)

For these reasons, the contrasting groups, borderline group and borderline regression methods which are all criterion based approaches, were purposefully selected for this study as they are appropriate for establishing decisions of competence in high stakes situations such as promotion or certification. These three methodologies were also chosen for several other reasons. Firstly, they are examinee based, which means that the data used to establish passing scores is generated from actual examinee performance and collected at the same time of the exam; making them feasible and efficient, eliminating further time commitments on the part of the examiners. Secondly, they need only basic statistical analyses making them accessible to exam administrators and educators. Lastly, modifications to these approaches have been described for
the purpose of performance based assessments, allowing for easy application of these methods to this study’s OSATS data. (J. J. Norcini, 2003)

The consistency of the passing scores across methodologies using both a compensatory and conjunctive model, provides support for any of these methods to be applied confidently to OSATS performance data giving educators options in their standards setting choice, and allowing for the application of the method that most suits their needs. The decision to use one method over the other can then be based on the capabilities of the assessment infrastructure, local expertise and size of exam. For example, one of the main disadvantages of the borderline group method is its use in small scale exams where there is a risk that the cut score is based on the performance of a relatively small number of individuals. (Wood et al., 2006) This is overcome with the borderline regression method, which uses all of the examinees’ data. Recognizing these types of limitations and knowing that the results across the three methods are similar will allow evaluators to choose a methodology with more confidence. Our results also demonstrate consistency across a compensatory and conjunctive model, again suggesting that both can be used keeping in mind the goal of the evaluation tool.

Interestingly, the consistency of the set pass scores seen in this study is not replicated in the OSCE literature, where different methodologies have demonstrated more variability. One explanation for the consistency seen in the present study is the large sample size: the calculations for the pass scores were derived from a much larger sample than is typically found in the OSCE literature. Wood et al. (2006) have acknowledged the instability in cut scores across different methods when the sample size is small. (Wood et al., 2006) A second explanation for the stability in cut scores is perhaps the more objective expectation of a ‘competent’ performance when an examiner is evaluating technical skill. It is conceivable that constructs evaluated in the OSCE, such as communication skill and history taking, may be displayed in various ways yet still achieve the same score. By contrast, there is rarely a great variability in the approach to a technical task; for example, there is little variability in performing a hand-sewn anastomosis that would be considered ‘competent’. The more objective conceptualization held by the examiners in terms of what a ‘competent” performance looks like may be another explanation for the consistency across the three methods. Finally, the stability of pass scores may be partly related
to the fact that all three approaches used in this study are examinee based, which means that cut scores are based on actual observations of examinee performance. Test based standards setting approaches, by contrast, are based on judgments about the exam content. The variability seen in some standard setting studies for the OSCE may be due in part to the comparison of these two different types of approaches. (Kramer et al., 2003; Livingston & Zieky, 1989; Schoonheim-Klein et al., 2009)

This is the first large scale study to set technical skills performance standards using several standard setting methodologies, comparing a compensatory and conjunctive model. The results of this study have major implications for developing and improving existing competency based technical skills curricula. Existing competency based programs describe arbitrary or unclear standards, allowing residents to progress in their competency based training once they have “…achieved preset benchmarks of what we believed represented technical competence”. (Ferguson et al., 2013) OSATS it is one of the best validated tools and is considered the ‘gold standard’ for assessing technical skill even for high stakes decisions of competence. (Martin et al., 1997; R. K. Reznick & MacRae, 2006; van Hove et al., 2010) It is an assessment tool that is widely applicable to educators and administrators across surgical specialties. This large, robust database has been used to apply, compare and assess standards setting methodologies and demonstrated that it is feasible to establish defensible and credible standards of performance. These results can provide guidance in the selection and implementation of these methodologies in current and future competency based programs.

Further study is required to explore whether a pass/fail status on the OSATS exam can predict future technical skill. If the predictive ability of the OSATS passing scores is established, both the OSATS exam and the passing scores set with this study will have a dramatic and meaningful effect on the early identification and remediation of the underperforming surgical trainee.

The major limitation of this study is its retrospective nature. Typically standard setting involves the prospective selection of a standard setting method, careful selection of judges (or as in this study, examiners), and collection of data. Although the data used for this study was retrospectively analyzed, the three standards setting methodologies were appropriate for the
dataset and methodologically straightforward to apply. Furthermore, inconsistency in the examiners over time was minimized by a standardized examiner training session and the understanding that candidates’ performance was to be evaluated regardless of their level of training.

4.6 Conclusion

Standard setting is an essential component of competency based training and assessment, which has been overlooked within the surgical assessment literature. This is the first large scale study to introduce and apply standard setting methodologies using the world’s largest OSATS database and the first to demonstrate both the feasibility and consistency of the three methodologies. These results can help guide and educate exam administrators in selecting and implementing the most feasible and appropriate standard setting methodology for their assessment tool.
5 Chapter 5: Passing a technical skill exam in the first year of residency predicts future performance

Accepted and In Press at the Journal of Graduate Medical Education

5.1 Abstract

Objective: To establish if passing the Objective Structured Assessment of Technical Skill (OSATS) exam predicts future performance.

Background Data: Predicting performance is the ultimate goal of assessment which could help with the early identification of residents at risk.

Design: Between 2002-2012, 133 general surgery PGY1 residents at the University of Toronto completed an eight station simulated OSATS exam as a mandatory component of their training program. With recently set passing scores, residents had been assigned an overall pass/fail status using three standards setting methods (contrasting groups, borderline group, borderline regression). Future in-training performance was then compared between residents that had passed and failed the OSATS using retrospectively collected in-training evaluation reports (ITERs) from resident files. A Mann-Whitney U test compared performance between groups at the PGY2 and PGY4 level.

Results: Residents that passed the OSATS exam outperformed those that failed, when compared in their PGY2 year across all 3 three standard setting methodologies (p<0.05). At the PGY 4 year the contrasting groups method continued to show a significant difference (p<0.05), while a trend continued with the borderline group and borderline regression methods.
**Conclusions:** This is the first study to predict resident performance using pass/fail results on a technical skills exam. Residents who pass the OSATS are more likely to outperform those who fail when compared in PGY2; with this trend continuing in PGY4. Predicting performance is the gold standard of assessment. The identification of underperforming residents is a key implication from these findings, allowing for the opportunity for early remediation.
5.2 Introduction

Competency based surgical education is gaining momentum around the world, aiming to ensure that surgeons throughout their training achieve the necessary skill to provide safe patient care. (Accreditation Council for Graduate Medical Education, 2015; Alman, Ferguson, Kraemer, Nousiainen, & Reznick, 2013; Sonnadara et al., 2014; The Royal College of Physician and Surgeons of Canada, 2014) Establishing competence at a point in time, however, does not ensure ongoing or future competence. While documenting competence is an essential component of competency based education, the gold standard of assessment is the ability to predict competence in the real world setting. (Southgate et al., 2001) Having the ability to predict future performance would have major implications on resident selection, promotion and certification.

While a surgeon is expected to achieve competence in several domains, technical skill remains a key component for surgical specialties. While training in a simulated environment has demonstrated transferability to the operating room,(Vanderbilt et al., 2015; Zendejas, Brydges, Hamstra, & Cook, 2013) using simulation to assess technical skill and predict future ability during residency has not been demonstrated. A number of tools have been developed to assess technical skill,(van Hove et al., 2010) but one of the most widely used platforms is the Objective Structured Assessment of Technical skill (OSATS) (Hatala et al., 2015) which has been implemented across a variety of specialties with minor adaptations to the originally described multi station format.(Argun et al., 2015; S. L. de Montbrun et al., 2013; B. A. Goff et al., 2002; Hatala et al., 2015) Although widely implemented and studied, one of the limitations of the original OSATS exam has been its lack of a set passing score, limiting its use in pass/fail decisions. (R. K. Reznick & MacRae, 2006) Furthermore, there are no data investigating the predictive ability of this exam. Recently, passing scores have been set for the original OSATS exam which has allowed individuals to be classified as competent or non-competent based on their pass/fail status.(S. de Montbrun, Satterthwaite, & Grantcharov, 2016)

One way to build evidence of validity for the OSATS exam and the recently set passing scores is to investigate the impact of the pass/fail status and the ability of the results of the OSATS to predict future performance in the real clinical environment (Cizek & Bunch, 2007; Hatala et al.,
2015) If the OSATS exam is predictive of future clinical technical skills performance, this would provide validity evidence for the set passing scores and have implications on surgical training, including the early recognition and remediation of underperforming trainees. To this end, the purpose of the present study was (1) to build evidence of validity for the recently set OSATS passing scores and (2) to investigate if passing or failing the OSATS exam predicts future technical performance during residency training.

5.3 Methods

This study was conducted at the University of Toronto and the research ethics board at St. Michael’s Hospital approved this study. The initial performance data was collected from the OSATS exam database. The OSATS exam is an 8 station simulation based technical skills examination that takes place at the University of Toronto’s Surgical Skills laboratory. All surgical residents (general surgery, plastic surgery, orthopaedic surgery, cardiac surgery, neurosurgery, otolaryngology) from 2002-2012 completed this examination at the end of their first year as a mandatory component of their core junior residency training (n=513). Surgeons from all divisions at the University of Toronto were recruited to serve as expert examiners. The examiners underwent a 30 min orientation and training session on the day of the examination to ensure the OSATS scoring system was understood. This examination is not repeated again during the 5 year training program.

A recent study has set and compared passing scores for the OSATS exam using three different standard setting methodologies; contrasting groups method (CG), the borderline group method (BG), borderline regression method (BR). A description of these three standards setting methodologies and their use in setting passing scores for the OSATS is described in detail elsewhere. (S. de Montbrun, Satterthwaite, et al., 2016) This study demonstrated that all three methodologies can be used to set pass/fail scores effectively and comparably for the OSATS exam.

The present study used OSATS pass/ fail status for the general surgery resident group only (n=133) as compared to all surgical specialties (n=513). During the 10 year study period
typically 12-13 residents were enrolled per year in the University of Toronto’s general surgery program.

Until recently, the OSATS did not have a passing score and residents were only assigned a raw score; upon completion of the OSATS, the raw score was given to both the resident and their program director. However, the recently set passing scores, allowed the 133 general surgery residents to be retrospectively assigned a pass/fail status according to all three standard setting methodologies. (S. de Montbrun, Satterthwaite, et al., 2016) This pass/fail status data were used for the present study; residents were not informed of their pass/fail status. A compensatory model was used. Therefore residents were given a “pass” status if they met or surpassed the overall OSATS passing score or a “fail” status if they did not meet the passing score.

The pass rate for the 133 general surgery residents was 79.7 per cent (106 of 133), 84.2 per cent (112 of 133) and 83.5 per cent (111 of 133) for the contrasting groups, borderline group and borderline regression methods respectively. (S. de Montbrun, Satterthwaite, et al., 2016)

The present study then compared future residency performance between the passing group and the failing group using all three standards setting methodologies to evaluate the predictive ability of passing or failing the OSATS during PGY1.

Future performance was assessed using in-training evaluation reports (ITER). PGY2 and PGY4 ITER data were retrospectively reviewed and collected from the general surgery residents’ training files. The general surgery site coordinators completed the ITERs at each training hospital by gathering input from all the surgeons that worked with the resident. The site-coordinators were selected because they have an interest in residency training and experience with the ITER scoring system. These individuals did not have access to the trainees’ first year OSATS exam score.

The ITERs include multiple domains of competence, however, only data specific to technical skill were retrieved. Throughout the study period the ITER form changed. Prior to 2004, the
ITER forms contained three items reflecting technical skill 1) performance of technical procedures 2) overall technical performance and 3) intra-operative decision-making. These three items were each evaluated on a 5 point Likert scale, with descriptors anchoring points for scores 1, 3 and 5. In 2004, the ITER form was updated to include 6 items adapted from the original OSATS global rating scale. These items were 1) knowledge of procedure 2) tissue handling 3) time and motion 4) technically competent for level of training 5) intra-operative judgment and 6) overall competence. These items were also rated on a 5 point Likert scale. A composite overall technical skill score for each resident was assigned by calculating a mean score out of 5 combining all items from their ITERs for their PGY2 and for their PGY4 year. This standardized score allowed for consistency in comparing individuals who may have had a different number of evaluations, and for comparison across the two different ITER forms.

A Kolmogorov-Smirnov test of normality was conducted on the overall technical skill score for the PGY2 and PGY4 data. A Mann Whitney U test was used to compare the overall technical skill score at PGY2 and at PGY4 between residents who passed or failed the OSATS exam using all three standard setting methodologies.

5.4 Results

Of all the general surgery residents that took the OSATS exam during the study period (n=133), ITER data were available to calculate an overall technical skill score on 109 PGY2 residents and 76 PGY4 residents. The Kolmogorov-Smirnov test of normality demonstrated a deviation from normal ($p<0.05$) on the overall technical skill score for PGY2s and PGY4s, therefore the non-parametric Mann-Whitney U test was used to compare the groups.

The majority of PGY2s (n=63, 57.8%) had data from 2 ITERs (range 1-3) which were used to calculate their overall technical skill score. The majority of PGY4s (n=63, 82.9%) had 2 or 3 ITERs (range 1-4) which were used to calculate their overall technical skill score.
At the PGY2 level, a statistically significant difference in overall technical skill score was seen between residents that passed and failed the OSATS according to all three standard setting methods, (CG, BG, BR) with those who passed outperforming those who failed (Mann Whitney U CG $z = 3.49, p < 0.001$; BG $z = 2.50, p = 0.01$; BR+ $z = 2.09, p = 0.04$) (Table 9). At the PGY4 level, this statistically significant difference was still present using the CG method, with passing residents continuing to demonstrate a higher overall technical skill score than failing residents (Mann Whitney U $z = 2.58, p = 0.01$) (Table 10). While a trend in performance was seen with the BG and BR methods in the PGY4 this difference was not statistically significant (Mann-Whitney U BG $z = 1.37, p = 0.17$; BR $z = 1.17, p = 0.24$) (Figure 14).

**Table 9.** Comparison of PGY2 overall technical skill score between residents that passed and failed the OSATS

<table>
<thead>
<tr>
<th>Standard Setting Methodology</th>
<th>PGY2 Overall Technical Skill Score</th>
<th>OSATS fail (Median, IQR‡)</th>
<th>OSATS pass (Median, IQR‡)</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrasting Groups</td>
<td></td>
<td>3.83 (0.75) (n= 23)</td>
<td>4.22 (0.50) (n= 86)</td>
<td>$p &lt;0.001$</td>
</tr>
<tr>
<td>Borderline Group</td>
<td></td>
<td>3.89 (0.76) (n= 17)</td>
<td>4.19 (0.49) (n= 92)</td>
<td>$p = 0.01$</td>
</tr>
<tr>
<td>Borderline Regression</td>
<td></td>
<td>3.94 (0.84) (n= 18)</td>
<td>4.17 (0.44) (n= 91)</td>
<td>$p = 0.04$</td>
</tr>
</tbody>
</table>

‡IQR – Inter quartile range
5.5 Discussion

Establishing the achievement of competence is paramount for surgical training. However, the ability to predict future performance is the ultimate goal of assessment. The ability to identify residents who will underperform would have major implications for surgical training, allowing for the early identification of individuals that could benefit from structured remediation and deliberate practice. This study demonstrates that the pass/fail status on the OSATS exam during the first year of general surgery training has the ability to predict future technical performance. Our results show that general surgery residents who fail the OSATS exam according to all three standard setting methods are more likely to underperform based on ITER overall technical skill scores, in their PGY2 year. As time passes the ability to predict performance becomes more

Table 10. Comparison of PGY4 overall technical skill score between residents that passed and failed the OSATS

<table>
<thead>
<tr>
<th>Standard Setting Methodology</th>
<th>PGY4 Overall Technical Skill Score</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OSATS fail (Median, IQR‡)</td>
<td>OSATS pass (Median, IQR‡)</td>
</tr>
<tr>
<td>Contrasting Groups</td>
<td>3.67 (1.29)</td>
<td>4.10 (0.71)</td>
</tr>
<tr>
<td></td>
<td>(n=20)</td>
<td>(n=56)</td>
</tr>
<tr>
<td>Borderline Group</td>
<td>3.72 (1.39)</td>
<td>4.04 (0.73)</td>
</tr>
<tr>
<td></td>
<td>(n=14)</td>
<td>(n=62)</td>
</tr>
<tr>
<td>Borderline Regression</td>
<td>3.78 (1.33)</td>
<td>4.00 (0.74)</td>
</tr>
<tr>
<td></td>
<td>(n=15)</td>
<td>(n=61)</td>
</tr>
</tbody>
</table>

‡IQR – Inter quartile range
difficult as more variables influence outcomes; despite this, the contrasting groups methodology continued to predict performance and show a statistically significant difference between groups even at the PGY4 level. The loss of statistical significance (despite a clear trend in performance differences) in the PGY4 year for the borderline group and borderline regression methods, does not discount them as useful and credible standard setting methods, rather this study highlights the limitation of distant prediction and the need for continuous assessment throughout training.

It is not uncommon to see surgical residents who have successfully progressed through their training despite evaluators feeling that the individual is underperforming. Despite evaluators’ subjective ability to recognize poor performance, it is often not reflected in students’ final evaluations. (Cohen, Blumberg, Ryan, & Sullivan, 1993b) This has been referred to as the failure to fail phenomenon. Lack of reliable systems for summative assessment, remediation options, anticipation of an appeal process and lack of documentation are potential explanations for this phenomenon. (N. L. Dudek, M. B. Marks, & G. Regehr, 2005) The responsibility of failing a resident is often left to subsequent supervisors, where the cycle of failing to fail typically continues with underperforming residents continuing to pass. Progression within a general surgery program often relies on ITER evaluations which are known to be poor at identifying residents with below average technical skill. (Feldman et al., 2004) Implementing an objective assessment of technical skill early on in surgical training that has the ability to predict future performance, would be instrumental in identifying underperformers, and introducing early educational interventions for effective remediation. Evidence on the reliability and predictive validity of these objective tests will allow the implementation of curricula for summative assessment and will help to address the failure to fail phenomenon. The ability to predict underperformers would also help to counteract the tremendous economic and psychological burden suffered by the resident, the training program and society.

Previous reports in the literature have suggested that ITERs are poor at identifying residents with below average technical skill. (Feldman et al., 2004) In contrast, our results suggest that ITER scores may be able to discriminate resident technical performance. We found that a failing score on the OSATS in the PGY1 year was associated with significantly poorer overall technical skill scores on the PGY2 ITER. This difference was maintained in the PGY4 year when data were
analyzed using contrasting groups methodology. The absolute difference in subsequent median ITER scores, however, was small between groups. The median ITER scores for those participants who failed the PGY1 OSATS ranged from 3.83 to 3.94 in the PGY2 year and from 3.67 to 3.78 in the PGY4 year. In contrast, the median ITER scores for those who passed the OSATS ranged from 4.17 to 4.22 in the PGY2 year and 4.00 to 4.10 in the PGY4 year. This suggests that a score of 3 (the mid-point of the scale) with a descriptor of competent, may be overestimating performance at that level. This rightward shift of the assessment scale is consistent with the existing literature that ITER data is typically heavily biased towards competent. Despite this bias, the present
**Figure 14.** Boxplots comparing overall technical skill score at the PGY2 and PGY4 level between residents who pass and fail the OSATS exam during PGY1 using a. contrasting groups method b. borderline group method c. borderline regression method
study was still able to show a difference in ITER score between groups. Given that ITER evaluations are already well established in many training programs, it is important to recognize this upward shift when interpreting an individual resident’s ITER data.

In contrast to the ITER, the OSATS has accrued a wealth of validity evidence for the interpretation of its scores. (Martin et al., 1997; R. K. Reznick & MacRae, 2006; van Hove et al., 2010) However, its use in high stakes decisions has been limited due to the lack of an established pass score. (R. K. Reznick & MacRae, 2006) Setting pass scores and investigating the impact of pass/fail status addresses the ‘implication or decisions’ component of Kane’s model of validity. (M. Kane, 2013) This domain of validity has been neglected in the OSATS validation literature, and is an essential component if the OSATS is to be considered for high stakes assessment decisions. (Hatala et al., 2015) However, recently, this domain of validity has been addressed and pass scores for the OSATS have been set and compared using several standard setting methods. (S. de Montbrun, Satterthwaite, et al., 2016) Until recently, only a few studies have addressed the issue of pass/fail scores for an OSATS type of exam, typically with a pass/fail decision based on overall dichotomous pass/fail judgment reflecting a candidates’ overall performance, rather than using known structured, rigorous standard setting methodologies. (B. Goff et al., 2005; B. A. Goff et al., 2001; B. A. Goff et al., 2002) Moreover, no study to date has looked at the implications of passing or failing an OSATS exam. (Hatala et al., 2015) The present study aimed to build on the ‘implication or decisions’ validity argument by looking at the predictive ability of the pass/fail status on future residency performance. Demonstrating that residents who pass the OSATS outperform in their future surgical training when compared to those who fail builds validity evidence for the OSATS exam, as well as validity evidence for the recently set pass scores. Furthermore, the concept of using an OSATS type of structured assessment of technical skill has been studies by the American Society of Colon and Rectal Surgeons. Recognizing that technical skill is an essential component of surgical training that should be formally evaluated upon completion of training, the ASCRS has developed and studied the implementation of a similar type of assessment tool. (S. L. de Montbrun et al., 2013) Building evidence of multiple sources of validity for these types of assessments is essential as the surgical community continues to move towards incorporating these performance based skills exams.
The use of technical skill simulation to assess and then predict future performance in the real clinical environment is a relatively new concept. Traditionally, simulation training has been used as an adjunct to teach fundamental technical skills outside of the operating room thus flattening the learning curve inside the real operating room. Many studies have demonstrated beneficial transfer of skills acquired in the laboratory to the operating room. (Vanderbilt et al., 2015; Zendejas et al., 2013) However, data on using simulation to assess and predict future performance are limited. Moore et al (2014) used a simulated technical skill assessment during residency selection to predict future performance during residency, demonstrating a moderate correlation. A weakness of this study was that the authors did not use a dichotomous pass/fail status, limiting the ability to identify the cohort of individuals who are at risk of future difficulties. While both studies have successfully demonstrated the used of simulation to predict performance in the real clinical environment, the advantage of the present study is its ability to dichotomize the group into a competent and non-competent cohort with an evidence-based cutoff score and thus identify underperforming residents that would benefit from early remediation.

The OSATS pass/fail scores are a new finding and have not yet to be implemented into practice at the University of Toronto. Questions remain in terms of what to do with residents that “fail”, and how to best provide remediation or support for these residents. Further work will need to be done to study remedial programs and the process or re-testing post remediation.

This study has three primary limitations. One limitation of the present study is the use of ITER data to evaluate future in training technical skill performance. ITER data, as already mentioned, have been criticized for being unreliable, and poor at identifying below average residents with regards to technical skill. (Feldman et al., 2004) Inconsistency has also been seen when comparing written comments on the ITERs with the numerical ITER score, with the written comments suggesting a lower level of performance than the numerical score. (Richards, Campbell, Walshaw, Dickens, & Greco, 2009) However, while the reliability of the ITER can be very low with a single rater, aggregated ITER data, as used in the present study, with multiple evaluators have been shown to be more reliable. (Carline, Paauw, Thiede, & Ramsey, 1992; Ginsburg, Eva, & Regehr, 2013; Littlefield, Paukert, & Schoolfield, 2001) The present study addressed the unreliable nature of ITER data by collecting and aggregating resident ITER scores,
and recognizing that the ITER data were completed by different evaluators. Another limitation related to the collected ITER data, was the switch in ITER form in 2004. The first collection of PGY2 ITER data would have been in 2003, meaning that only approximately 1 year of ITER data and only approximately 1 cohort of residents would have used the old evaluation form. Moreover, despite the different terminology, the Likert scales of the two ITERs were essentially the same allowing for the calculation of a mean score of all items regardless of the ITER used.

The second of this study is its retrospective nature from a single institution. Although a limitation, it allowed for the collection of a decade of performance data resulting in a relatively large sample size not typically found in the education literature. The retrospective nature also limited the completeness of the initial assessment as well as the follow-up data, with 18.0% and 42.9% loss to follow-up in the PGY2 and PGY5 year respectively. Potential reasons for this included: 1) residents changing programs 2) incomplete resident files 3) clerical error with ITERs misfiled or lost and 4) for the PGY4 year, the last cohort did not yet have PGY4 data available at the time of the study. While a statistically significant difference was seen in the PGY2 year, the more substantial loss to follow-up seen in the PGY4 year may have partly contributed to the inability to show a difference between groups. Lastly, the study is limited to a single institution and a single surgical specialty, which may limit its generalizability to other surgical specialties and educational environments. Future study should aim at collecting prospective data on specific operative assessments of common procedures in general surgery and to initiate a multi-center implementation and evaluation of a summative assessment program.

5.6 Conclusion

The present study is the first to demonstrate the ability of the OSATS pass/fail status to predict future technical performance of general surgery trainees. This work also addresses the current lack of evidence surrounding the use of technical skills examinations for summative assessment. The combination of previous evidence using standard setting methodologies to set a pass/fail scores for the OSATS and the present findings demonstrating the predictive ability of the pass/fail status, suggest that the OSATS could be used in decisions of promotion or resident
remediation. A key implication of these findings is the early identification and remediation of the underperforming resident. Future, multi-center work investigating the predictive ability of assessments throughout residency selection, certification and beyond would be of great value for surgical education and patient safety.
Chapter 6: Implementing and evaluating a national certification technical skills exam: the Colorectal Objective Structured Assessment of Technical Skill (COSATS)

This chapter has been modified from the following:


6.1 Abstract

**Objective**: To implement the Colorectal Objective Structured Assessment of Technical skill (COSATS) into American Board of Colon and Rectal Surgery (ABCRS) certification and build evidence of validity for the interpretation of the scores of this high stakes assessment tool.

**Background Data**: Currently, technical skill assessment is not a formal component of board certification. With the technical demands of surgical specialties, documenting competence in technical skill at the time of certification with a valid tool is ideal.

**Methods**: In September 2014, the COSATS was a mandatory component of ABCRS certification. Seventy candidates took the exam, with their performance evaluated by expert colorectal surgeons using a task specific checklist, global rating scale and overall performance scale. Passing scores were set and compared using two standard setting methodologies, using a compensatory and conjunctive model. Inter-rater reliability and the reliability of the pass/ fail decision were calculated using Cronbach’s alpha and Subkoviak’s methodology respectively. Overall COSATS scores and pass/fail status were compared to results on the ABCRS oral exam.

**Results**: The pass rate ranged from 85.7% - 90%. Inter-rater reliability (0.85) and reliability of the pass/fail decision (0.87 and 0.84) were high. A low positive correlation ($r =0.25$) was seen
between the COSATS and oral exam. All individuals who failed the COSATS passed the ABCRS oral exam.

**Conclusion:** COSATS is the first technical skill exam used in national surgical board certification. This study suggests that the current certification process may be failing to identify individuals who have demonstrated technical deficiencies on this standardized assessment tool.
6.2 Introduction

Assessing and documenting the achievement of competence is a major focus for surgical education around the globe, with surgical boards and regulatory bodies working towards developing and implementing documentation of milestone progression. (Cogbill et al., 2014; The Royal College of Physician and Surgeons of Canada, 2013a) Although, each transition in surgical training is important, one could argue that the transition from trainee to independent surgeon is particularly important, as it assumes readiness for unsupervised practice. While assessment of knowledge and judgment occurs at the time of board certification, internationally, technical skill typically continues to be assessed with case logs, in-training evaluation reports, and procedural based assessments, with no formal or direct assessment of technical skill at the time of certification. Commonly used methods of assessing technical skill lack the reliability and validity that is necessary for high stakes assessment, such as certification. (Feldman et al., 2004; Sidhu, Grober, Musselman, & Reznick, 2004) Given the importance of technical competence in the care of surgical patients, with evidence that improved technical skill leads to improved outcomes, (Birkmeyer et al., 2013) formally and objectively documenting technical skill should be a mandatory component of certification decisions.

The colorectal objective structured assessment of technical skill (COSATS) is a technical skill exam that was developed in 2012, to assess competence in colorectal technical skill at the time of certification. (S. L. de Montbrun et al., 2013) Candidates rotate through 8 technical skill tasks in the surgical skills lab and their performance is evaluated by expert examiners. The original pilot study from 2013, demonstrated good interstation reliability, and some evidence of validity. (S. L. de Montbrun et al., 2013) However, because the COSATS exam is being developed for the purpose of certification, an extensive body of validity evidence is required to support the interpretation of COSATS scores. (Korndorffer et al., 2010) Messick’s framework of validity was used to guide the validity evidence sought in the present study. (Messick, 1989) In September of 2014, the COSATS exam was made a mandatory component of American Board of Colon and Rectal Surgery (ABCRS) certification exam, allowing for data on the entire examinee cohort to be analyzed in this validation study.
With the COSATS exam being developed as a high-stakes certification technical skill exam, the purpose of this study was to build evidence of validity for the interpretations of the COSATS scores. Using Messick’s framework, three sources of validity were sought (1) Internal structure evidence – including inter-rater reliability and the reliability of the passing score (2) Relationship to other variables evidence – looking at the relationship between the COSATS and ABCRS oral examination results and (3) Consequences evidence – setting a credible passing score for the COSATS using standard setting methodologies and evaluating the pass/fail rate.

6.3 Methods

The research ethics board at St. Michael’s Hospital in Toronto approved this study.

6.3.1 Structure of the COSATS Exam

The COSATS exam has been previously described. (S. L. de Montbrun et al., 2013) The present study included 5 of the original models, and the development of 3 new task simulations. The operative tasks selected for the COSATS were based on expert consensus after reviewing the objectives of colon and rectal surgery training and the ABCRS procedure list. (American Board of Colon and Rectal Surgery, 2011) New simulated models were pilot tested by 2 board certified colorectal surgeons who were not involved in the study, to ensure feasibility of the task in the allotted time as well as the ability of the model to represent the intended operative task.

6.3.2 Assessment Instruments

At each station candidates were evaluated using (1) a task specific checklist (2) a global rating scale (GRS) and (3) an ‘overall skill’ scale. Task specific checklists were developed for the 3 new stations using an iterative process. The original COSATS global rating scale (GRS) was modified in order to add detailed descriptors at each point along the scale. The GRS was used at all stations and consisted of 6 dimensions of operative performance (Figure 15. COSATS global rating scale). Candidates were also evaluated on their overall performance at each station using an ‘overall skill’ scale which was developed to collect data for standard setting.
purposes. This consisted of a Likert scale which described in detail the performance characteristics and overall skill level for each of the five points on the scale (Figure 16), and provided an overall gestalt of performance for each station.

6.3.3 Participants

In September of 2014, the COSATS exam was a mandatory component of the ABCRS examination for all first time takers of the written and oral exam. Although mandatory to complete the COSATS for certification, the results did not impact certification, and a ‘no standing’ status was assigned.

6.3.4 Examiners

Board certified colorectal surgeons served as examiners. They were asked to review all 3 evaluation tools one week prior to the exam and also underwent a 30 minute training session the day of the examination, including a discussion around the expected performance of candidates at each point along the Likert scale.
<table>
<thead>
<tr>
<th>RESPECT FOR TISSUE</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inappropriate force during majority of procedure, resulting in tissue damage</td>
<td>Frequently uses unnecessary force on tissue or causes damage to tissue</td>
<td>Reasonably careful handling of tissue but some damage from inappropriate handling</td>
<td>Almost always handles tissue appropriately, little damage to tissue</td>
<td>Handles tissue with care throughout entire procedure, no damage to tissue</td>
<td></td>
</tr>
<tr>
<td>TIME AND MOTION</td>
<td>Many unnecessary moves</td>
<td>Frequent unnecessary moves, moves the case forward slowly, several unsure movements</td>
<td>Some unnecessary moves</td>
<td>Few unnecessary moves</td>
<td>No unnecessary moves</td>
</tr>
<tr>
<td>Not progressing or moving case forward</td>
<td></td>
<td></td>
<td>But still progresses and moves case forward</td>
<td>Good economy and efficiency</td>
<td>Maximal economy and efficiency</td>
</tr>
<tr>
<td>INSTRUMENT HANDLING</td>
<td>Very awkward and tentative handling and movement of instruments throughout procedure</td>
<td>Frequent awkward/ tentative movement</td>
<td>Minimal stiffness and appropriate handling of instruments</td>
<td>Fluid without awkward movements</td>
<td>Expert/ skilled fluidity of movement and handling of instruments</td>
</tr>
<tr>
<td>FLOW OF OPERATION</td>
<td>Unsure of next steps, significant hesitation moving forward</td>
<td>Some initial forward progression but stalls and fails to get to next steps</td>
<td>Slow but steady forward progression of case, minimal hesitation</td>
<td>Clear knowledge of next step, easily moves the case forward</td>
<td>Seamless movement to next step of procedure</td>
</tr>
<tr>
<td>EXPOSURE</td>
<td>Unable to optimize visualization and set up for task/ procedure for the majority of the case</td>
<td>Fumbling to gain good visualization, with compromised exposure and awkward tissue alignment</td>
<td>Exposure adequate with occasional compromise, but able to recover</td>
<td>Good exposure throughout the procedure</td>
<td>Optimal exposure throughout</td>
</tr>
<tr>
<td>(CONSIDER THE USE OF ASSISTANTS OR THE USE OF RETRACTORS)</td>
<td>Fumbling to gain good visualization, with compromised exposure and awkward tissue alignment</td>
<td>Exposure adequate with occasional compromise, but able to recover</td>
<td>Good exposure throughout the procedure</td>
<td>Good alignment of tissue for the majority of the time</td>
<td>Clear visualization</td>
</tr>
<tr>
<td>KEY COMPONENTS OF PROCEDURE</td>
<td>Key component missed, Egregious error ,Unsafe maneuver</td>
<td>Key components done inadequately, or acknowledged but not completed</td>
<td>Key components completed adequately</td>
<td>Key components completed safely and with increased skill</td>
<td>Key components of the task are clearly recognized, acknowledged and executed skillfully</td>
</tr>
</tbody>
</table>

**Figure 15.** COSATS global rating scale
<table>
<thead>
<tr>
<th>Overall Skill</th>
<th>Unsatisfactory</th>
<th>Non-Competent</th>
<th>Adequately Competent</th>
<th>Clearly Competent</th>
<th>Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>This candidate clearly requires more training and was <strong>grossly inadequate</strong> in performing this task. Would not be comfortable with this candidate performing this procedure independently in <em>training</em>. Unsafe to operate independently. Below the level of a colorectal resident. This performance is a CLEAR “fail”, and <strong>WELL BELOW the passing score.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>This candidate demonstrated basic technical ability, but <strong>inadequately performed</strong> this task. Would not be comfortable with this candidate performing this procedure independently in practice. This candidate would require significant supervision. At the level of a colorectal resident who still requires some training. This performance is <strong>BELOW the passing score.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>This candidate is <strong>adequately competent</strong> to complete this task safely and independently. The candidate could perform this procedure independently in practice. This candidate’s performance would fall <strong>RIGHT ON THE BORDER of the “pass” score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>This candidate is <strong>clearly competent</strong> to complete this task independently and safely. There is room for this individual to develop further expertise in technical ability. Fully comfortable with this candidate performing this procedure independently and safely in practice. This performance is <strong>ABOVE the passing score</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>This candidate demonstrated <strong>exceptional performance</strong> of this procedure. This candidate is technically sophisticated in performing this procedure, I am fully comfortable with this candidate performing this procedure independently and safely in practice. Exemplary technical skill, with a performance <strong>WELL ABOVE the passing score.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 16.** COSATS overall performance scale
6.3.5 Exam Administration

The COSATS exam was administered over two days on the same weekend as the ABCRS oral exam. The exam consisted of 8 technical skills stations and 2 rest stations. Each station was preceded by a stem describing the required technical task. The total testing time was 2.5 hours (12 minutes for task completion and 3 minutes for model turn over). One administration of the exam had 2 raters per station allowing for the assessment of inter-rater reliability.

6.3.6 Validity Evidence

6.3.6.1 Internal Structure Evidence

Inter-rater reliability of the exam was calculated using a Cronbach’s alpha ($\alpha$). Reliability of the pass/fail decision was assessed by establishing the agreement co-efficient ($p_0$) using Subkoviak’s methodology. (Subkoviak, 1988)

6.3.6.2 Relationship to Other Variables Evidence

A scatterplot visually explored the relationship between the COSATS scores and the ABCRS scores. A Pearson correlation coefficient ($r$) was used to measure the relationship between COSATS scores and ABCRS oral exam scores. Furthermore, a 2X2 cross tabulation table was used to explore relationship between the binary pass/fail decision on the COSATS and the ABCRS oral examination.
6.3.6.3 Consequences Evidence

The borderline group method and the borderline regression method were the two standard setting methodologies used to set passing scores for each COSATS station. An overall exam passing score was then calculated for both methodologies as the sum of the 8 station passing scores.

6.3.6.3.1 Borderline Group Methodology

The borderline group method sets the passing score at level of a ‘borderline candidate’. (Livingston & Ziesky, 1982) For the purpose of this study the ‘borderline candidate’ was an individual who was described as ‘adequately competent’; able to complete a task safely and independently and reflected a score of 3 on the ‘overall skill’ scale. For each station, all individuals that scored a 3 on the ‘overall skill’ scale represented this borderline group. The passing score for each station was then calculated as the mean of the global rating scale scores for this group. (Newble, 2004)

6.3.6.3.2 Borderline Regression Methodology

The borderline regression methodology uses a linear regression model to regress the global rating scale scores of all examinees onto their ‘overall skill’ score. An ‘overall skill’ score of 3 (the performance of a borderline candidate) is inserted into the linear equation to determine the predicted global rating passing score for each station.

6.3.6.3.3 Pass Criteria for the COSATS

In addition to setting passing scores for each COSATS station and the overall exam using a borderline group and borderline regression methodology, two different models were used to establish an individuals’ overall examination pass/fail status. Passing criteria can be set using a compensatory or a conjunctive model. A compensatory model allows an examinee to pass the COSATS by achieving the overall exam passing score, even if they fail some stations (allowing high scores on some stations to compensate for low scores on others). (Newble, 2004) In contrast,
a conjunctive model requires that candidates achieve both the overall exam passing score, and pass a certain number of stations. A compensatory and conjunctive model were compared in order to evaluate consistency across the two models; with a conjunctive model candidates were required to pass based on overall exam score in addition to passing at least 50% of the stations.

6.4 Results

A total of 70 candidates who were concurrently taking their ABCRS exam took the COSATS exam and 34 colorectal surgeons served as COSATS examiners.

6.4.1 Internal Structure Evidence

The inter-rater reliability ($r$) of the COSATS was high at 0.85. The reliability of the pass/fail decision (agreement co-efficient) ($p_o$) was also very high at 0.87 for the borderline group methodology and 0.84 for the borderline regression method. This reliability is in the range necessary for high stakes assessment, such as certification.

6.4.2 Relationship to other Variables Evidence

A scatterplot of COSATS exam score and ABCRS oral exam score suggests that the two assessments are evaluating different constructs, with a low positive correlation ($r=0.25$) (Figure 17). The comparison of the binary pass/fail status between the COSATS exam and the ABCRS oral exam is outlined in Table 11, Table 12, Table 13 and Table 14.

They all demonstrate that with both standard setting methodologies and with both a compensatory and conjunctive model that, every candidate that failed the COSATS exam passed the ABCRS oral exam.
Figure 17. Scatterplot comparing COSATS score with the ABCRS oral exam score
Table 11. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a compensatory borderline group method

<table>
<thead>
<tr>
<th></th>
<th>COSATS Fail (n)</th>
<th>COSATS Pass (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Exam Fail (n)</td>
<td>0 (0%)</td>
<td>7 (10%)</td>
</tr>
<tr>
<td>Oral Exam Pass (n)</td>
<td>7 (10%)</td>
<td>56 (80%)</td>
</tr>
<tr>
<td>Total</td>
<td>7 (10%)</td>
<td>63 (90%)</td>
</tr>
</tbody>
</table>

Table 12. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a conjunctive borderline group method

<table>
<thead>
<tr>
<th></th>
<th>COSATS Fail (n)</th>
<th>COSATS Pass (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Exam Fail (n)</td>
<td>0 (0%)</td>
<td>7 (10%)</td>
</tr>
<tr>
<td>Oral Exam Pass (n)</td>
<td>9 (12.9%)</td>
<td>54 (77.1%)</td>
</tr>
<tr>
<td>Total</td>
<td>9 (12.9%)</td>
<td>61 (87.1%)</td>
</tr>
</tbody>
</table>
Table 13. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a compensatory borderline regression method

<table>
<thead>
<tr>
<th></th>
<th>COSATS Fail (n)</th>
<th>COSATS Pass (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Exam Fail (n)</td>
<td>0 (0%)</td>
<td>7 (10%)</td>
</tr>
<tr>
<td>Oral Exam Pass (n)</td>
<td>10 (14.3%)</td>
<td>53 (75.7%)</td>
</tr>
<tr>
<td>Total</td>
<td>10 (14.3%)</td>
<td>60 (85.7%)</td>
</tr>
</tbody>
</table>

Table 14. Comparing the pass/fail status of the ABCRS oral exam with the pass/fail status of the COSATS exam using a conjunctive borderline regression method

<table>
<thead>
<tr>
<th></th>
<th>COSATS Fail (n)</th>
<th>COSATS Pass (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Exam Fail (n)</td>
<td>0 (0%)</td>
<td>7 (10%)</td>
</tr>
<tr>
<td>Oral Exam Pass (n)</td>
<td>9 (12.9%)</td>
<td>54 (77.1%)</td>
</tr>
<tr>
<td>Total</td>
<td>9 (12.9%)</td>
<td>61 (87.1%)</td>
</tr>
</tbody>
</table>
6.4.3 Consequences Evidence

Passing scores and rates for each of the 8 COSATS stations, as well as overall exam passing score using the borderline group and borderline regression methodology are outlined in Table 15. Station passing scores and station passing rate were very similar across the two standard setting methodologies with little variation; the laparoscopic ileorectal anastomosis station had the largest difference in both station score and station pass rate. The overall exam passing score was also similar across the two standards setting methodologies; 147.8/240 (61.6%) and 150.4/240 (62.7%) for the borderline group method and borderline regression method respectively.

The overall exam pass rate for all methodologies was in the range expected for high stakes exams such as certification. The overall exam pass rate for the borderline group method was 90% and 87.1% using a compensatory and conjunctive model respectively. The overall exam pass rate for the borderline regression method was 85.7% and 87.1% using a compensatory and conjunctive model respectively.

The overall assigned binary pass/fail decision showed very little variation between the two standard setting methodologies; there was also little variation in assigned pass/fail status between a compensatory and conjunctive model (Table 11, Table 12, Table 13 and Table 14).
Table 15. COSATS technical skill stations with mean scores, passing scores and passing rates

<table>
<thead>
<tr>
<th>Station</th>
<th>Mean score (SD) ‡ (n=70)</th>
<th>BG pass score (SD)</th>
<th>BG pass rate</th>
<th>BR pass score</th>
<th>BR pass rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectal Prolapse</td>
<td>19.3 (4.6)</td>
<td>18.2 (1.8)</td>
<td>57.1</td>
<td>18.7</td>
<td>57.1</td>
</tr>
<tr>
<td>Pelvic bleed</td>
<td>21.1 (3.6)</td>
<td>18.7 (1.7)</td>
<td>77.1</td>
<td>18.9</td>
<td>77.1</td>
</tr>
<tr>
<td>IPAA</td>
<td>21.3 (4.3)</td>
<td>18.7 (1.8)</td>
<td>72.9</td>
<td>18.8</td>
<td>72.9</td>
</tr>
<tr>
<td>CA anastomosis</td>
<td>22.7 (3.0)</td>
<td>18.9 (1.5)</td>
<td>88.6</td>
<td>19.0</td>
<td>88.6</td>
</tr>
<tr>
<td>Lap IR anastomosis</td>
<td>22.1 (3.0)</td>
<td>18.6 (1.3)</td>
<td>90.0</td>
<td>19.7</td>
<td>77.1</td>
</tr>
<tr>
<td>Colonoscopy</td>
<td>21.3 (3.5)</td>
<td>18.4 (1.7)</td>
<td>71.4</td>
<td>18.7</td>
<td>71.4</td>
</tr>
<tr>
<td>Handsewn anastomosis</td>
<td>19.7 (3.6)</td>
<td>18.5 (1.4)</td>
<td>54.3</td>
<td>18.5</td>
<td>54.3</td>
</tr>
<tr>
<td>Lap sigmoidectomy</td>
<td>19.2 (4.3)</td>
<td>17.8 (1.4)</td>
<td>62.9</td>
<td>18.1</td>
<td>54.3</td>
</tr>
<tr>
<td>OVERALL EXAM PASSING SCORE</td>
<td>147.8/240</td>
<td>(61.6%)</td>
<td>150.4/240</td>
<td>(62.7%)</td>
<td></td>
</tr>
</tbody>
</table>

‡ Maximum score of 30 for each station
SD – standard deviation, BG – borderline group methodology, BR borderline regression methodology, IPAA – ileal pouch anal anastomosis, CA – coloanal, IR – ileorectal
6.5 Discussion

Milestones have been defined as competency based outcomes that are progressively demonstrated at significant points throughout training and practice. (Education, 2015a) With the ACGME moving towards continuous accreditation, specialty groups have focused on developing milestones programs that address the six ACGME core competencies. (Accreditation Council for Graduate Medical Education, 2001) Technical skill, although certainly not the only aspect of surgical competence, is essential for safe surgical practice, and it has been shown that surgeons with lower surgical skill ratings have higher rates of complications, reoperation and readmission as well as increased mortality. (Birkmeyer et al., 2013) Ideally, competence in technical skill would become a mandatory component of surgical board certification, assuming the availability of reliable, valid and feasible testing platforms.

This study is the first time that a surgical board within North America has implemented a certification technical skill exam. The COSATS exam was originally described in 2013, and was developed in conjunction with the American Society of Colon and Rectal Surgeons (ASCRS) to address the gap in technical skill assessment at the time of certification. (S. L. de Montbrun et al., 2013) Although initial evidence of validity has been demonstrated, given the significance of the implications of the results of this high stakes exam (false positives resulting in passing someone that is incompetent, false negatives resulting in failing someone that is competent) (Downing, 2004) validity evidence needs to be continually sought to support or refute the interpretations of the COSATS scores. (Downing, 2003) The purpose of this study was to build evidence of validity for the interpretations of the COSATS scores using Messisk’s framework; this validation study sought evidence of internal structure, relationship to other variables and consequences evidence.

Our findings suggest that the COSATS is feasible to incorporate into the certification process and has a level of reliability that is acceptable for high stakes decisions. The reproducibility of the pass/fail decision is a very important aspect of validity evidence for certification, ensuring
that individuals would receive the same status across multiple iterations of the exam. (Gugiu, Gugiu, & R.) Both the agreement co-efficient and the inter-rater reliability of the COSATS was over 0.8, the level typically necessary for high stakes exams. (Downing, 2003)

A recent review, recognized the need for further validity evidence for the use of OSATS type of examinations in the setting of high stakes assessment. (Hatala et al., 2015) Our study aimed to address this by using and comparing rigorous and structured standard setting methodologies to establish a passing criteria for the COSATS. Presently the literature on standard setting for OSATS type of examinations has been based on assigning an overall dichotomous pass/fail judgment reflecting a candidates’ overall performance, rather than the implementing structured standard setting methods. (B. Goff et al., 2005; B. A. Goff et al., 2001; B. A. Goff et al., 2002) This study is the first study to apply and compare rigorous standard setting methodologies, using both a compensatory and conjunctive model, for an OSATS structured examination. The pass/fail status assigned with this study allowed for the calculation of the overall exam passing rate to assess the impact of the set scores. It also allowed for the comparison of pass/fail status between the COSATS and the oral ABCRS exam.

To evaluate if the COSATS exam was measuring something different than the current oral exam, we compared scores on the COSATS to scores on the ABCRS oral exam. We found a low positive correlation, suggesting that the COSATS is measuring a unique construct and if added to the certification process would increase the validity of the board exam as a whole. A key finding in this study was evident when the pass/fail status on the COSATS exam was compared to the pass/fail status on the ABCRS exam. We found that despite the standards setting methodology used, or the use of a compensatory/ conjunctive model, that all individuals who failed the COSATS technical skill examination passed the ABCRS oral examination, and have been certified. This suggests that perhaps adding an assessment of technical skills might help in making certification decisions.

Although, tools to assess technical skill have been developed over the past few decades, only two have been described and implemented into high stakes decisions; the Fundamentals of
Laparoscopic Surgery (FLS) and the European Board of Vascular Surgery (EBVS) technical skill exam.

The technical skill component of the FLS program is based on the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) (Vassiliou et al., 2006) and assesses basic laparoscopic skills. This tool has established evidence of validity and has set pass/fail scores allowing decisions of competence in basic laparoscopic skill to be made (Fraser et al., 2003). The FLS program is not part of the American Board of Surgery (ABS) certification exam in general surgery per se; however, since 2009, it has become a mandatory pre-requisite to the exam. (Surgery, 2012) The major limitation of FLS is that it assesses very basic tasks, such as knot tying and PEG transfer without the ability to assess advanced technical skill required for certification in general surgery. The COSATS exam has been developed to assess a much broader domain of technical skill, and is designed to assess advanced open and laparoscopic procedures and tasks. The COSATS aims to evaluate the requisite skill within the broad domain of colon and rectal surgery, making it more appropriate for technical skill assessment of specialty surgical certification.

In the early 2000’s the EBVS developed a technical skill exam for vascular surgery board certification. The exam consists of three vascular surgical tasks on bench-top models and the initial pilot study described construct validity, inter-rater reliability and internal consistency of the exam. (Pandey et al., 2006) Since 2004, this technical skill exam has been incorporated into the EBVS certification process. One serious threat to the reliability and validity of this exam is case specificity, (Downing, 2004) which occurs when the exam consists of only a few tasks which are intended to represent competence in a broad domain. Too few tasks may not adequately represent true ability in the broader domain. More scenarios mitigate the overall impact of content specificity giving a more reliable test. The COSATS exam has used eight technical skill stations to overcome this potential limitation, as it has been previously shown that eight stations gives a reliable indicator of performance. (Martin et al., 1997) Secondly, with the EBVS exam, a rigorous methodology to set a passing score was not described, with an arbitrarily set passing score of 75%, without evidence of validity of this passing score. The present study, alternatively, used and compared two different, methodologically sound standard setting methodologies, and also compared a compensatory and conjunctive model. The consistency seen
across all methods, demonstrates initial evidence of credibility of the COSATS passing scores. The EBVS has not published any further data on the ongoing development or evaluation of their certification tool.

Several key issues related to the future of the COSATS continue to be discussed amongst the ABCRS and the ASCRS. Timing of the administration of the COSATS has been debated. Administering the COSATS in conjunction with the ABCRS oral exam makes sense from a logistics perspective; however, it means that we are testing individuals who are already practicing surgeons. Assessing technical skill may be more appropriate prior to entering practice, near the end of colorectal training, as this would allow for the opportunity to remediate while still enrolled in a surgical training program. The future of the COSATS is currently uncertain, with ongoing discussion about timing of the exam, remediation of failing candidates, and gathering further psychometric evidence on the exam.

Secondly, the authors recognize that the COSATS exam has limitations with its simulated models; advanced technical dissection skills required for colorectal surgery such as rectal mobilization, or pouch lengthening maneuvers are not easily replicated. The COSATS exam represents basic colorectal technical skill that a colorectal surgeon in practice should be able to perform. We would expect that most individuals would have little difficulty in passing this exam, as is seen in the present study. However, the individuals who are underperforming are demonstrating a lack of ability on basic technical tasks, highlighting the need to identify these individuals. Lastly, the process of board certification already requires a financial commitment on the part of the candidate, and although a formal cost analysis has yet to be conducted, it has been estimated that the COSATS cost per candidate would be approximately $1,000. Prior to full incorporation of the COSATS into training or as an ABCRS requirement these issues need to be resolved.

6.6 Conclusion

As the surgical community continues to move towards competency based assessment, reliable tools will be needed. The COSATS exam has demonstrated a level of reliability and evidence of validity necessary for high stakes assessment and is one example of how to assess colorectal
technical skill at the time of certification. The COSATS is the first technical skill exam that has been successfully incorporated into national surgical board certification. While the future of the COSATS remains uncertain, this study suggests that the current certification process may be failing to identify individuals who have demonstrated technical deficiencies on a standardized assessment tool. Implementing a reliable and valid technical skill exam as a component of certification, will help to ensure that newly trained surgeons have the necessary skill to provide safe patient care.
7 Chapter 7: General Discussion and Limitations

Parts of this chapter will be published in the textbook entitled: Simulation for Surgery and Surgical Subspecialties. Springer. Chapter 11; Use of Simulation in High-Stakes Summative Assessments in Surgery.

7.1 Thesis Summary

While the last decade has demonstrated the development of technical skills assessment platforms, few have been developed for the purpose of high stakes assessment. Recognizing this gap in surgical assessment research, the present thesis has focused on key issues in developing a high stakes examination. Developing a high stakes exam requires a high level of validity evidence for the interpretation of the scores. The projects within this dissertation address several key stages in the development and implementation of high stakes technical skills assessments using a contemporary validity framework. The specific aims were defined in order to highlight and address existing gaps in knowledge within the domain of high stakes technical skills exam development.

Each paper within this dissertation addresses a key step in the evolution of the development of a high stakes exam from the initial construction of an exam blueprint, to the implementation of a technical skills exam into board certification, to the setting of a passing score and finally an exploration of the consequences of the examination’s pass/fail status.

In the Specific Aim 1, the first key step in developing a high stakes examination was addressed, through the development of a blueprint for a technical skills examination for graduating general surgery residents in Canada. A Delphi methodology was successful at gaining consensus amongst Canadian general surgery residency program directors. Content related validity evidence was generated from this study which resulted in a detailed list of tasks and procedures that a graduating Canadian general surgery resident should be able to perform independently in practice. The methodology used in this paper can serve as a guide to the development of
examination blueprints for any surgical specialty. The results of this Delphi serve as a solid foundation for the construction of simulated laboratory based bench top models and the selection of items to be included in the final examination.

In the Specific Aim 2, we sought to address a major gap in the surgical assessment literature. Namely, the introduction, application and comparison of standard setting methodologies as they apply to performance based technical skills data. Three standard setting methodologies were purposefully selected for this study. Using the world’s largest OSATS database, with OSATS performance data from 513 PGY1 surgical residents, these standard setting methodologies were applied in order to generate passing scores for the OSATS exam. The results of the passing scores and the pass/ fail status were compared across the three methodologies. Interestingly, all three methods produced very stable results in terms of passing scores and pass/ fail status, suggesting that all three methodologies can be applied to this type of performance data, and the selection of method based on the properties of the exam and local expertise.

In the Specific Aim 3, was to evaluate the predictive ability of the pass/fail status. The gold standard of assessment is the ability to predict performance. To date, resident performance has not been predicted using results from an objective assessment of technical skill. The final manuscript in this dissertation demonstrated that general surgery residents who had passed the OSATS exam (based on the set passing scores from Specific Aim 2) during their PGY1 year had significantly better technical skill in their PGY2 year than those general surgery residents that failed the OSATS in their PGY1 year, based on in-training evaluation report data on technical skill. This trend continued to be seen in the PGY4 year with those that had passed the OSATS continuing to outperform those that had failed. This is a very interesting finding, as no previous study has predicted performance based on pass/fail status of a technical skills exam. The predictive ability of the OSATS pass/fail status certainly stimulates questions regarding the usefulness of the widespread implementation of this examination early on in surgical training to help identify residents at risk and initiate structured remedial curriculum.

The final Specific Aim 4 a technical skills examination for colorectal surgery (the Colorectal Objective Structured Assessment of Technical Skill (COSATS) exam was implemented as the
first North American technical skills certification examination. The results are provocative and demonstrated that individuals who fail the COSATS technical skills examination, pass their written and oral exam, suggesting that the COSATS is identifying individuals with technical deficiencies, that with the current certification process, would otherwise go on to become American Board of Colon and Rectal Surgery certified. This study highlights the importance of the inclusion of a technical skills exam into surgical certification.

Over the past many decades, the acquisition and assessment of resident technical skills has taken place in the operating room. The operating room as a platform to teach and assess skill has been questioned in response to issues of time, cost, patient safety issues and ethical considerations of “testing” on real patients.

Over the past few decades the concept of the surgical skills laboratory had gained immense interest around the world. The surgical skills laboratory provides a safe environment to learn and test surgical skill. (R. K. Reznick & MacRae, 2006) The laboratory also provides the opportunity for high stakes standardized testing. Simulation is an ideal platform for high stakes assessment of technical skill.

### 7.2 The Use of a Unified Contemporary Model of Validity in Surgical Education

The use of a unified contemporary model of validity in the simulation education literature is spares, and has typically focused on few sources of validity. A recent systematic review by Cook et al. investigating the use of validity theory in simulation based education, found that many studies fail to outline the type of validity evidence sought with their study design and even fewer interpreted the results of their findings within a contemporary theory of validity.(Cook et al., 2013)

Within the domain of surgical education, a recent review found similar results, demonstrating a considerable lag in the adoption of contemporary validity frameworks, despite the advancements of validity within parallel assessment fields.(Ghaderi et al., 2015)
The surgical literature has also focused mostly on two main validity measures, leading to the suggestion that current assessment measures be used solely for the purpose of instruction or formative assessment. For an assessment to be used in high stakes assessment, validity evidence needs to be extensive. (Korndorffer et al., 2010)

Others have also acknowledged the lack of “consequences” evidence within Messick’s framework, also suggesting that current tools not be used for high stakes assessment. Consequences evidence has been recognized as an area requiring further investigation. Most specifically within the OSATS literature, consequences evidence and the domain of validity addressing implications and outcomes from the examination have not been studied. (Hatala et al., 2015)

The studies presented in this dissertation have addressed this major gap in validity theory as it applies to performance based assessments in technical skill. Each study has used a unified, contemporary validity theory. Each study has also described, at the outset, the type of validity evidence to be sought and has also interpreted the results within the validity framework. Furthermore, the studies have broadened the type of validity evidence that is captured in the literature. The consequences evidence that has been lacking in the surgical literature was thoroughly investigated within the studies of this thesis. The use of this unified and contemporary validity framework within the present work has been recognized by educational experts in the field; as recently stated by Arora and Darzi:

“...very few studies have utilized appropriate standard setting approaches and that setting pass marks arbitrarily, although the current norm, is poor practice; the current article advances the science of assessment by moving beyond this. In addition, despite initial testing of the COSATS tool in a previous study, the authors are to be commended on their attempts to generate further validity evidence using a contemporary theory of validity. They have recognized that validation evidence is never complete. Promoting the ongoing evaluation of COSATS in this way has strengthened its robustness and credibility and providing an example to other researchers of how additional evidence can, and should be gathered to support the inferences made from test scores.”(Arora & Darzi, 2016)
7.3 Simulated High Stakes Assessment Contextualized in Miller’s Theory of Assessment

Miller’s pyramid serves as a useful taxonomy for understanding assessment, with each level describing assessment methods at progressive skill levels (Figure 1). This theory allows for the selection of assessment methods to match the skill being evaluated. (Miller, 1990)

This pinnacle level of assessment within Miller’s taxonomy of assessment represents the “Does”, and assesses skill in a real world setting with real patients. This is the most authentic uncued professional assessment. An example of assessment at the “does” level would be the observation and evaluation of a resident’s performance during a surgical case such as a right hemicolectomy.

The peak of the pyramid in Miller’s theory represents the highest level of assessment. Intuitively, it would make sense to consider this type of assessment the most appropriate for high stakes assessments such as surgical residency promotion or surgical certification. While this level would represent the most valid in terms of real world performance, it has several disadvantages for its use in technical skills assessment.

The “shows how” level of assessment is the second highest level in Miller’s Pyramid. This level aims at assessing if students are able to demonstrate or “show” what they “know how” to do. (Miller, 1990) This level of assessment typically involves performance based assessments; this is the level of assessment that is addressed with the OSATS, GOSATS and COSATS simulated technical skills examinations. While these assessments may be somewhat artificial, they elevate the assessment beyond just knowledge.

The rational and argument for the use of simulated performance based assessments in the evaluation of technical skill is outlined in the following section.
7.4 The Rational for Using Miller’s “Shows How” Level of Assessment for High Stakes Assessment of Technical Skill

There are several advantages of using simulation for both the training and assessment of surgical skills, which will be reviewed in this section. This type of assessment falls within Miller’s pyramid as a “shows how” level of assessment (Figure 1).

Firstly, and of utmost importance is the safety and ethical issues that are avoided with simulation. Simulators allow residents to develop and learn skills through deliberate practice without impacting the health of a real patient. A simulated surgical scenario can allow room for error, allowing for completion of a task and thereby, allow the consequences of an error to be fully played out. This is very important since it allows the resident to realize the consequences of error. (Maran & Glavin, 2003)

Secondly, simulators can be used repeatedly and are reproducible, and since they do not rely on human beings, simulators can be standardized. Any assessment or learning on real patients must take into account significant patient variability. For example, a right hemicolectomy in a morbidly obese patient, who has had previous abdominal surgery is a completely different operation than a right hemicolectomy in a virgin abdomen, thin patient. Since so much variability can be found between patients when it comes to operative intervention, standardization is key in proving an equal “playing field” for residents to demonstrate their technical skill. Simulation provides this opportunity for standardization. The variability in patients and operative cases is avoided with simulation, making the assessment fair for all candidates. Standardization also allows for the development of reproducible platforms for the purpose of testing and formal assessment, making simulation ideal for summative testing purposes. While testing performance in the real world setting on real patients may be the most valid assessment when considering pyramid of assessment, this is less applicable for high stakes technical skills assessment for the aforementioned reasons. The assessment of other non-technical domains of competence required of physicians, such as communication, professionalism or judgment is not as influenced by patient body habitus, and thus more
amenable to workplace based assessment. Furthermore, the consequences of poor performance are less significant in these other domains of competence, whereas poor technical performance tested in the context of real clinical can lead to significant patient morbidity or even mortality. These reasons support the use of simulation in technical skills assessments.

Lastly, simulated platforms are available at any time to be used. (Issenberg et al., 1999) Patients are real human beings that are available to the learner during very specified times. Also, OR time is scheduled rigorously so again operative exposure occurs during specified times. Simulators are much more available and flexible than real patients allowing surgical programs and licensing bodies to incorporate simulation into resident training and resident licensing examinations.

7.5 Fluidity of Competence

Competence has recently been defined as an individual who possesses a minimum standard of performance to provide safe and independent surgical care. (Szasz et al., 2014) The technical skills examinations outlined in this dissertation attempt to establishing competence in technical skill at various levels of training. What needs to be considered is the concept that ‘competence’ is a fluid phenomenon. This notion of falling out of ‘expert’ status has been described, whereby experts may fall into a lesser category of expertise becoming an ‘experienced non-expert’. (Moulton, Regehr, Mylopoulos, & MacRae, 2007) The same holds true for competence. While the OSATS, GOSATS and COSATS aim to establishing competence at various points in training, they do not necessarily predict future performance. Chapter 5 tries to address this point by investigating the ability of the OSATS to predict future performance. While the results of this study suggest that we have the ability to predict future performance, further work needs to be done to evaluate if the COSATS has the ability to predict actual performance in practice. This is an essential component of future investigation as predicting performance is the gold standard of assessment; future work will address this issue.
7.6 Implementation of High Stakes Assessments for Training and Board Certification

Implementing high stakes performance based assessments into both training as a component of promotion and into board examination as a component of certification will present a major challenge in terms of feasibility (cost and size of exam), dealing with failing candidates (appeal process and development of remedial programs) and concerns of ‘training’ to the test.

Decisions regarding the future of the COSATS are currently being addressed by the executive council of the American Board of Colon and Rectal Surgeons. The COSATS is likely to become a mandatory component of ABCRS certification process. However, several issues need to be addressed with this possibility. Fortunately, the Medical Council of Canada has experience with the study and implementation of a performance based assessment of clinical skills into a high stakes licensing examination. Lessons learned from the implementation of the OSCE into the LMCC examination provide some guidance for the implementation of technical skills assessments in the domain of surgery.

7.6.1 Feasibility

No matter how reliable or “valid” an assessment is deemed to be, a major obstacle to implementation is feasibility of administration. Both the OSATS and COSATS exams are extremely labor intensive in terms of model development, exam administration, examiner time commitment, the cost of exam set up, models and lab rental.

The feasibility of this type of simulated examination is certainly a major obstacle to overcome. The first question that needs to be addressed by the surgical community at large is the importance of testing and documenting the achievement of technical competence of surgeons. Formally documenting technical skill both during and at the time of certification, would assure regulatory bodies as well as the public that surgeons have met the minimum standard of performance required for training or practice. One must weigh the importance of assessing and
documenting technical skill, with the feasibility, cost and labor involved in developing and administering this type of performance based assessment.

7.6.1.1 Cost of Taking the COSATS Examination

While a formal cost analysis had not been carried out on the COSATS exam, it has been estimated that the cost would be approximately $1,000.00 per candidate. When compared to the current cost of the Canadian General Surgery Board exam (which is well over $1,000.00)(The Royal College of Physicians and Surgeons of Canada, 2014c) and the American Board of Colon and Rectal Surgery certification exam (which is written $700 for the written examination and $800 for the oral examination)(The American Board of Colon and Rectal Surgery, 2012) the cost of a technical skills exam does not seem unreasonable.

When the MCC implemented the LMCC exam, a graded implementation of cost was used. It has been suggested that for the first several years while the COSATS is mandatory, but while candidates receive a designation of ‘no standing’, candidates could be charged a nominal fee for the exam. This would introduce a cost associated with the COSATS exam and will be fair to all subsequent cohort where a fee will be inevitable. For the first 3 years candidates could be significantly undercharged, following this, the cost of the COSATS could increase to allow for funds for ongoing test development and subsequent cost recovery.

7.6.1.2 Large Scale Exam Administration

Another issue of feasibility is the size of the exam. With the COSATS, implementing the exam for the 70 candidates demonstrated to be feasible. However, on an annual basis, approximately 100 candidates take the American Board of Colon and Rectal Surgery examination. Thus, the COSATS exam would have to accommodate 100 candidates.

As the broader surgical community gains interest in this type of objective assessment of technical skill, feasibility on a larger scale needs to be considered. Based on the knowledge from this PhD the American College of Surgeons has approached to collaborate on developing a technical skills
examination for senior general surgery residents in the United States. Considering issues of feasibility, one needs to consider the number of general surgery residents in the United States. For the academic year of 2015-2016, as reported by the Accreditation Council for Graduate Medical Education (ACGME), there were 262 accredited general surgery programs in the United States, with a total of 7907 residents on duty. (Accreditation Council for Graduate Medical Education, 2016) Assuming a 5 year general surgery program, this would be on average a total of 1581 residents per year. A technical skills exam, such as the GOSATS administered to this number of residents annually certainly will pose logistical challenges.

One way to address issues of exam size and feasibility is to develop an examination infrastructure. This would involve training surgical skills labs across the United States and Canada to be able to administer this type of performance based assessment.

Throughout North American surgical skills centers have the opportunity to become an American College of Surgeons accredited educational institute. These ACS accredited surgical skills labs will have the opportunity to become examination sites for the COSATS (and GOSATS) exam. Decentralization of the examination is a major objective for the near future. Building a testing infrastructure will involve developing a training curriculum for the involved ACS accredited skills laboratories.

This curriculum will include three components and will include 1) a didactic manual 2) video instructions on model building and 3) a site visit from an expert. The didactic manual will be a “How To” manual that will address issues in planning, setting up and administering a performance based assessment of technical skill. The video component will include instruction on the construction of each technical skills station. This will describe the task description, the materials needed for model development, a step by step description on building the models, and an outline of the surgical equipment needed for the station. Lastly, a site visit will take place, whereby, an expert will review the models and go over planning for an exam administration in term of laboratory set up and training. The labs will have the opportunity through this process to become an official examination test centre.
7.6.2 Dealing With the Failing Residents

One question related to the outcomes of a high stakes exam is what to do with residents that fail the exam. How will these individuals be remediated? Will these individuals require re-testing? How many opportunities will they be allowed to re-test before they are required to retrain? Developing and evaluating remedial strategies will be required component of the development of a testing program. While this issue is one that is already faced by regulatory bodies in terms of dealing with residents that fail the written or oral examination of surgical board examinations, this issue seems to become most heated when discussing technical skill or surgeons.

In presenting the results of this thesis, questions regarding remediation of technical skill are often raised. Certainly, a pathway for dealing with a failing resident will need to be clearly outlined. Remedial strategies will need to be developed to deal with these individuals.

One issue with the current administration of the COSATS is the timing of the examination. Currently, the COSATS has been administered in conjunction with the oral board exam. The reason for administering the exam in conjunction with the oral board exam related to feasibility. The oral exam provided the opportunity to capture the entire cohort of candidates in one location, allowing the COSATS to be administered in one city over one weekend. However, currently the ABCRS oral examination takes place when candidates have already been out in independent colorectal surgical practice for one year. This means that individuals taking this technical skills exam are already in independent practice. The thought of identifying technical deficiencies in surgeons who are already in practice creates a level of anxiety amongst regulatory bodies, boards and the public. Ideally, one would want to identify deficiencies in technical skill prior to entering practice.

There is a plan to change the timing of the COSATS exam such that candidates take the exam while they are still in colorectal surgery training. The recommendation would be administer the exam during training. Since colorectal surgical training is only one year, ideally administering the exam at approximately 7-8 months into the academic year would allow for the opportunity to
identify struggling residents, allow time for remediation and allow for the opportunity to
demonstrate correction of deficiencies. The timing of the exam remains to be firmly established,
keeping these issues in mind.

Because this is a high stakes examination, there will also need to be a formal appeal process in
place to address candidates who challenge their exam results.

### 7.6.2.1 The Appeal Process

Another major issue that will need to be addressed is the creation of an appeal process. The
following paragraph will outline a proposed appeal process that could be implemented for the
COSATS examination.

The only grounds for appeal are administrative problems. Anything regarding the administration
of the exam that a candidate feels either disadvantaged them or advantaged others (i.e. breakdown of a station model). The candidate will be informed that a successful appeal will not necessarily result in a “pass” standing, but that after full review of the appeal it is possible that a “no standing” status is assigned with the requirement that the candidate be reassessed and achieve a “pass” at the expense of the ABCRS.

The appeal process will include a three step process: 1) a telephone call from the candidate 2) a
written request to the board of governors of the American Board of Colon and Rectal Surgeons
from the candidate for an appeal 3) a review of the candidate’s grounds for appeal at a board
council meeting.

The first step in the appeal process involves a telephone call to the board, where individuals
wishing to pursue an appeal can get personalized general feedback on their overall performance.
This provides candidates with general information on their weaknesses and areas to focus on for
further preparation. Individuals will be given personalized feedback and guidance for self-
assessment.
The second step will require a written request for an appeal from the candidate outlining in a formal letter the reasons and circumstances for the appeal.

The third step, will involve reviewing the candidate’s case and their formal letter of appeal at the board’s council meeting. The candidate will be provided with the opportunity to be in attendance at the meeting to present their case. The council will have the opportunity to debate the case and finalize their decision. Possible outcomes of the appeal include 1) pass standing 2) fail standing 3) no standing. If a ‘no standing’ status is decided, the candidate will be reassessed at the following administration of the examination at the expense of the board.

A special circumstance may arise whereby a candidate demonstrates an egregious error or lapse in professionalism despite successfully passing the examination. These individual would need to be flagged and addressed. These candidates will be sent a letter from the board stating the concern with the observed behavior and be encouraged by the board to address this issue in their continuing professional development. Furthermore, the board would have the right to fail a candidate despite a passing mark if the candidate demonstrates egregious or unprofessional behaviour.

7.6.3 Exam Security and Training to the Test

The security of high stakes examinations is important. It would be impossible to prevent the distribution of examination questions and stations of the COSATS or GOSATS exam. With ongoing examination development and testing new tasks and models will be developed. While exam security is important, literature regarding the impact of exam content knowledge on examinee performance is quite interesting. The OSCE literature suggests that prior knowledge of content may not provide an advantage to candidates as one may expect.(Parks et al., 2006) Furthermore, it has been shown that prior knowledge of exam content can actually disadvantage a candidate. Swartz et. al conducted a study that was meant to deliberately violate test security by having the first group taking an OSCE exam provide detailed information to a second group who subsequently took the same OSCE exam. Despite this deliberate assault on test security the authors found that having access to information on exam content had little effect on exam
performance. (Swartz, Colliver, Cohen, & Barrows, 1993) When we consider technical skills assessment, for example, this is likely because if you do not know how to create an ileal pouch it is unlikely that knowledge of this task being on the exam would improve performance.

Training to the test is another concept to consider when designing this type of performance based assessment. Surgical programs may start to train candidates to the content of the examination. However, if the blueprint of the examination represents the broad range of technical skill required of that specialty, then putting more effort into training technical skills reflected in the exam would likely improve the learning experience of trainees. Furthermore over time, a larger bank of stations would be developed with a sample of tasks being chosen for any specific exam administration. Since exam content would change over time and since the content reflects a wide range of different procedures, exposing residents to these cases would only help to improve clinical training.

Furthermore, surgical regulatory bodies have specific requirements of training that candidates will still be required to achieve. (Surgery, 2012; The Royal College of Physician and Surgeons of Canada, 2013b) Thus, even if institutions train to test, they must still ensure that candidates are familiar and have completed a full and complete range of surgical procedures during their training.

### 7.7 Limitations

The use of simulation based technology to assess technical skill of surgeons or surgeons in training has several limitations. One of the biggest challenges is the acceptance of this method of assessment by stakeholders, regulatory bodies and surgeons themselves. As it stands currently, in general, the surgical community supports the assessment of knowledge and judgment with the present written and oral board examinations. However, as we have moved forward with the development of simulated technical skills exams there has been some push back from regulatory bodies and surgical boards to implement these assessments into high stakes assessment.
Change is difficult, and trying to advance the current board certification process to include technical skills assessment has proven challenging. Achieving buy-in from stakeholders represents perhaps the most significant challenge in the move towards high stakes technical skills assessment.

In order to achieve support from surgical boards and regulatory bodies a large and ongoing body of validity evidence needs to be accrued and ongoing study needs to be carried out. Furthermore, the belief as a surgical community in the importance of formal technical skills assessment needs to grow.

Implementing this type of technical skills assessment will be impossible without the full support from major regulatory bodies and surgical boards. The challenge will be to disseminate this knowledge in order to educate surgeons, surgeons in training, regulatory bodies and surgical boards about the importance of this type of assessment and the validity evidence to supports its use in high stakes decisions.

### 7.7.1 The Use of Simulation in High Stakes Assessment

The “shows how” level of assessment described by Miller may not represent the most valid and authentic approach to evaluating technical skill. The pinnacle of the pyramid is the “does” level, which assesses individuals in their actual clinical environment.

Assessing at the “shows how” level of Miller’s pyramid represents a limitation of the present studies, as it does not assess individuals in their clinical environment. However, for the reasons mentioned above, simulation was specifically chosen as the platform to assess technical skill in these high stakes situations because of its many advantages.

Firstly, simulation allows for models to be reproducible allowing all candidates to be exposed to a specific scenario regardless of actual incidence of that pathology in the real world setting. Secondly, simulated models can be standardized creating an equal playing field for all candidates. Lastly, simulation eliminates the ethical dilemma of “testing” technical skills on real patients, and thus avoids patient harm.
Although, the “does” level of assessment is perhaps the most valid as it tests individuals in the real world setting, simulation has many significant advantages for high stakes assessment such as certification.

7.7.2 Fluidity of Competence

Competence is a fluid phenomenon and deeming a candidate competent in a simulated setting at a specific point in time may not translate to competence in the real world setting and furthermore, may not translate into future competence.

One of the limitations of the present studies is recognizing that competence is a fluid phenomenon and establishing the achievement of competence at a specific point in time may not predict future performance.

The ability to predict future performance is the gold standard of assessment; the OSATS, GOSATS and COSATS are aimed at establishing competence at a point in time and while Chapter 5 demonstrates initial evidence in the ability to predict performance, further study into the predictive ability of the GOSATS and COSATS is essential.

Future study will attempt to evaluate how competence on these performance based assessments translates to both real clinical situations and how it translates to future performance in practice.

7.7.3 Simulated Model Limitations

Ideally simulated models would be able to replicate advanced procedures and tasks. However, simulated models do have limitations in their ability to fully replicate the complexities of human anatomy and anatomical planes.
The OSATS, GOSATS and COSATS models are limited in their ability to replicate procedures and tasks. For example, advanced technical dissection skills required for colorectal surgery such as rectal mobilization, or pouch lengthening maneuvers are not easily replicated. Furthermore, advanced general surgical procedures for the GOSATS, such as colonic mobilization and laparoscopic gastrectomy are also very difficult to simulate in the laboratory with bench top models.

Both the GOSATS and the COSATS exam represent basic general surgical and colorectal technical skill respectively, that a surgeon in practice should be able to perform. One would expect that most individuals would have little difficulty in passing the GOSATS and COSATS exam.
Chapter 8: Future Directions

Parts of this chapter will be published in the textbook entitled: Simulation for Surgery and Surgical Subspecialties. Springer. Chapter 11; Use of Simulation in High-Stakes Summative Assessments in Surgery.

This thesis has focused on the development high stakes technical skills exams for surgical residents. The PhD studies have attempted to build evidence of validity for the interpretation of test scores at various key stages in developing an examination. Each study has also attempted to address different domains of contemporary validity theory. While the evidence is enticing, the results have not surprisingly generated a plethora of further questions and areas of future study.

8.1 Knowledge Translation Plan

Knowledge translation and the dissemination of the results of this thesis is a major focus of future directions. The following measures will outline the knowledge translation plan to disseminate the knowledge gained from this PhD to the surgical community at large at a local, national and international level.

At a local level, the Program Director of the general surgery program at the University of Toronto has been informed through face to face discussions of the ability of the PGY1 pass/fail status on the OSATS to predict future in training performance. There are plans to apply the newly set passing scores to the OSATS exam. First year residents that fail the OSATS will be flagged as at risk residents. Remedial strategies will be considered for those individuals deemed at risk.

At an international level, results from both the study to set study setting passing scores for the OSATS exam and the study predicting future performance based in the pass/fail status have been presented at the American College of Surgeons annual meeting in 2014 and 2015.
respectively. Being incorporated into the educational session at the annual American College of Surgeons meeting allowed the results to target surgical education leaders and surgeons with an interest in surgical education.

The results of the Colorectal Objective Structured Assessment of Technical Skill have generated national and international interest. The COSATS work has been presented as an invited talk at the European Society of Coloproctology annual meeting in Dublin, Ireland in September 2015. The results of the COSATS have also been presented in an invited face to face meeting in Chicago with the Director of the Division of Education of the American College of Surgeons. From this work, and the results of our GOSATS Delphi, the American College of Surgeons has asked for partnership and collaboration in the development of a technical skills examination for US senior general surgery residents.

The COSATS has generated a dialogue within the surgical education community, with a recent editorial by Arora and Darzi published in the Annals of Surgery stating:

“*This article represents a landmark study which exemplifies how to use a contemporary theory of validity to gather psychometric evidence for an assessment tool. Rigorous standard setting methodologies are promoted and crucially, the tool is used to evaluate surgical skills at national board level. This is an important first step towards closing the implementation gap between educational research and practice.*” (Arora & Darzi, 2016)

The results of the COSATS have resulted in an invitation to speak as a guest panel speaker at the American College of Surgeons meeting 2016, in Washington, DC, with a focus on the process of developing and implanting a technical skills examination into board certification.

The COSATS has clearly generated international interest, and will pave the way for high stakes assessments in other procedural based specialties.
8.2 Assessing Technical Skill beyond Training and Certification

While this thesis has focused on the development, evaluation and implementation of performance based assessments into training and certification, the concept of testing technical skill should be considered throughout the continuum of a surgeons’ career.

Future projects will expand the target audience to include developing assessment platforms to assess surgeons in practice, assess the ageing surgeon and assess the technical skill of foreign trained surgeons who have an interest in credentialing within North America.

Furthermore, this dissertation has had a North American focus. The future will hopefully involve a more global approach, applying knowledge from this thesis to assessing surgeon skill around the world.

Developing ways to assess technical competence of new surgical innovation will also be necessary. For example, the introduction of TaTME as a new surgical approach (Atallah, Albert, & Monson, 2016) for low rectal cancer would benefit from an assessment tool to evaluate a surgeon’s competence in this area prior to implementation into their practice. Work in this area is underway.

The knowledge from this thesis provides the opportunity for other procedure based specialties to follow suit in developing objective assessments of technical skill, with the ultimate goal of improving patient outcomes.

8.3 Assessing Broader Domains of Surgical Competence

Technical skill is only one domain of competence required of surgeons. Certainly, surgeons need to achieve competence in other domains, such as knowledge, judgment and professionalism. This thesis has focused specifically on technical skill assessment, and the tools developed herein have remained focused on this domain of competence. Future endeavors will investigate the development of platforms to assess other domains of surgical competence. The knowledge gained from the developmental process, the implementation and the standard setting
for the high stakes assessment described in this dissertation will guide the development of platforms to assess various other domains of competence.

8.4 Developing and Implementing a Formative Assessment Program

While the current theme of surgical training is competency based, with a focus on assessing defined competencies at specified training points, there is a period of time between these assessments which provides the perfect opportunity to implement a formative feedback system. The present thesis has focused on specific time points throughout the spectrum of training. The OSATS has been used as an assessment at the time of transition from PGY1 to PGY2, the GOSATS is being developed as an assessment for the transition of senior residents from PGY5 to practice, and finally the COSATS has been developed for the purpose of certification after completing colorectal fellowship training.

While these assessments provide the opportunity to document the achievement of competence at specific time points (i.e. residency promotion, certification), the ongoing assessment of skill between these isolated snapshots of performance is lacking. Furthermore, these summative types of assessments unfortunately do not address the learning needs of the residents. The lack of a structured formative assessment program has been identified as an area of further study.

A project is currently underway to develop and implement a formative assessment system for general surgery residents at the University of Toronto. The goal of formative assessment is to provide students with on-going meaningful feedback, to motivate and encourage learning.

The objective of this pilot study is to investigate the development and implementation of a formative feedback system of daily resident operative performance using an electronic version of an already validated global rating scale. (Martin et al., 1997)
The benefits of a formative assessment system would be far reaching. This type of formative evaluation system could be implemented in any surgical specialty, would require little in terms of resources and would provide significant benefit to resident learning. Another major advantage would be the potential for early identification of residents in difficulty. Far too often residency programs are struggling with residents that are clearly underperforming, despite continuous passing evaluations. This “failure to fail” phenomenon (Nancy L Dudek, Meridith B Marks, & Glenn Regehr, 2005) is evident when supervisors, despite their ability to recognize poor performance, do not reflect this poor performance in students’ final evaluations. (Cohen, Blumberg, Ryan, & Sullivan, 1993a) Reasons for this include factors such as lack of documentation, lack of knowledge as to what to specifically document, the lack of remediation options, the lack of a record documenting the trainee’s day to day performance and the anticipation of an appeal process. A formative feedback system could help address these assessment issues.

The timing is ideal for the implementation of a structured formative feedback system for general surgery residency programs. The current focus on milestones, while important, has unfortunately left behind the ripe opportunity for rich formative assessment. This proposed system will fill the space between the milestones with the formative feedback needed for learning and improvement.
References


10.1016/j.jamcollsurg.2003.08.020


Objectives of Training in the specialty of General Surgery. (2010). Retrieved from Ottawa, Canada:


10.1067/msy.2000.108115


