The Operating Room as a Classroom: a New Approach to Skills Assessment and Training in Surgery

by

Esther Maria Bonrath

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

Institute of Medical Science
University of Toronto

© Copyright by Esther Maria Bonrath 2015
Abstract

Introduction: Ensuring that trainees are graduated as competent, safe surgeons who can independently perform successful operations is a social obligation. Much of surgical technical training occurs in the operating room; as a result, enhancing training in this environment needs to be a priority. In this thesis, a comprehensive training approach, based on coaching principles, is described as a means to structure the experiential learning experience of training in the operating room.

Methods: Three main steps were required. Step 1: Design of an assessment methodology based on error analysis, as an adjunct to global skill assessment, to allow for comprehensive feedback. Step 2: Standardization of training content. For this, delineation of erroneous from adequate task execution was required, and causal relationships between errors and intraoperative injuries (events) needed identifying. Step 3: Design of a coaching intervention, based on observed operative performance, using established instructional approaches.
**Results:** (1) The generic error rating tool was based on four error modes attributed to nine laparoscopic tasks. Substantial construct validity evidence was gathered through applying the tool in video analyses of routine procedures. (2) In an international Delphi consensus, definitions and examples of errors and events were obtained. Further, evidence supporting the role of errors was obtained through reviewing 66 error-event patterns identified by video analysis of routine bariatric procedures. (3) The effectiveness of the coaching intervention was demonstrated in a randomized controlled design. Coached residents (n=9) scored significantly higher on a procedure-specific skill scale (BOSATS) and made fewer technical errors than conventionally trained residents (n=9) (BOSATS median 3.90 (i.q.r. 3.68-4.30) vs. 3.60, (2.98-3.70), \(P=0.017\); technical error 10 (7-13) vs. 18 (13-21), \(P=0.003\)).

**Conclusion:** Designing training in a personalized fashion, following principles of coaching, can help structure experiential learning in the operating room, which leads to significantly enhanced skill acquisition in surgical training.
Acknowledgments

I would like to thank Dr. Teodor Grantcharov for his continued guidance and mentorship. Beyond being a fantastic supervisor, his exceptional patience and understanding were a source of reassurance during the complex processes surrounding all aspects of degree completion, without which I would have faltered in many occasions.

I would also like to extend my gratitude to my thesis committee members, Dr. Glen Bandiera and Dr. Avery Nathens. Their questions and comments prompted me to review, reassess, and improve numerous aspects of my research proposals. It has been a privilege to have the support of these two outstanding researchers throughout the last few years.

Research is a team effort, and I had the pleasure of working with a fantastic team. I especially wish to thank Dr. Boris Zevin, an inspirational role model of optimism, for the support, collaboration and friendship over the last years. Also I would like to thank Dr. Andras Fecso, Dr. Heinrich Husslein, Dr. Marisa Louridas, Dr. Eliane Shore, and Dr. Peter Szasz, all of whom contributed to the team spirit in the lab, offering friendship, and always open for stimulating discussions.

I would like to thank Karthik Raj and Dr. Jennifer Hickey for coordinating the numerous research and administrative processes during the past years.

Also, I would like to thank the information specialists at St. Michael’s Hospital’s health library, especially Ekaterina Lolova, Carolyn Ziegler and Teruke Kishibe, for their support during the development of literature search strategies throughout this research.

This research would not have been possible without the support of surgeons at the University of Toronto who supplied their cases for video review, or the residents that participated and invested their time during the coaching study. Therefore, I wish to thank all participants for their interest and commitment.
In addition, I wish to acknowledge the members of the expert panel that lent their expertise to the consensus process during the Delphi study. These are Dr. H.J. Bonjer, Dr. N. Christou, Dr. L.S. Feldman, Dr. P. Funch Jensen, Dr. M. Gagner, Dr. H. Lönroth, Dr. J. Mamazza, Dr. M. Morino, Dr. T. Olbers, Dr. J. Paige, Dr. R. Rosenthal, Dr. M. Schijven, Dr. C. Schlachta, Dr. N. Soper, Dr D. Stefanidis, Dr. S. Tsuda, Dr. M. Utech, Dr. B. Wolfe, Dr. T.M. Young-Fadok, as well as those who wished to remain anonymous.

Lastly, I owe my deepest gratitude to my husband Dr. Nicolas Dedy for his continuous support. As a researcher, team player, and co-author he guided me during all steps of this degree. His critical thinking helped me identify potential pitfalls early enough to avoid them during planning and conceptualization of projects. As my partner in life, his love and guidance kept me on track even through the darkest of times.
Contributions

I, Esther M. Bonrath, solely prepared this dissertation thesis and am the first author of all manuscripts arising from the research projects associated with this degree and detailed within this thesis. In all projects I designed the study, conducted the data collection and evaluation, interpreted the results, and drafted the final manuscript.

Dr. Teodor P. Grantcharov (Primary Supervisor and Thesis Committee Member), guided me throughout all phases of this degree as an academic mentor, lending his expertise to the design and conception of the projects, overseeing the data evaluation and interpretation as well as critical revision of the manuscripts arising from the research projects.

My Thesis Committee Members Dr. Avery Nathens and Dr. Glen Bandiera guided me throughout the studies offering valuable advice regarding planning and execution of the studies as well as interpretation of the results.

Dr. Nicolas J. Dedy, co-authored the manuscripts detailed in Chapters 3, 4 and 6 as well as the review detailed in Chapter 1, section 3.2.4. His role during these studies involved support in planning and execution of the studies, as well as analysis and interpretation of the results including critical revision of the final manuscripts.

Dr. Boris Zevin, co-authored the manuscripts detailed in Chapters 3 and 4 as well as the review detailed in Chapter 1, section 3.2.4. His role during these studies involved support in planning and execution of the studies, as well as analysis and interpretation of the results including critical revision of the final manuscripts.

Lauren Gordon, co-authored the manuscripts detailed in Chapters 5 and 6. Her role during these studies involved the programming of the analysis windows required for the video-analyses, technical support throughout the evaluation process, as well as interpretation of the results including critical revision of the final manuscripts.
Funding

Operating funds for the studies conducted and reported within this thesis were received from the University of Toronto, Education Development Fund (EDF), the Ontario Research Fund (ORF), and from Johnson and Johnson Medical Companies Canada.
Table of Contents

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

CONTRIBUTIONS ................................................................................................................ VI

FUNDING ................................................................................................................................ VII

TABLE OF CONTENTS ......................................................................................................... VIII

LIST OF TABLES .................................................................................................................. XII

LIST OF FIGURES ................................................................................................................ XIV

LIST OF APPENDICES .......................................................................................................... XV

LIST OF ABBREVIATIONS .................................................................................................... XVI

THESIS OVERVIEW ............................................................................................................. XVIII

1 LITERATURE REVIEW ....................................................................................................... 1

1.1 EXPERIENTIAL LEARNING THEORY ......................................................................... 2

1.2 MODELS OF POSTGRADUATE SURGICAL EDUCATION ............................................ 7

1.2.1 APPRENTICESHIP MODEL ....................................................................................... 8

1.2.2 EARLY RESIDENCIES AND THE ROLE OF PROFESSOR WILLIAM HALSTED AND PROFESSOR EDWARD CHURCHILL .................................................................................................................. 10

1.2.3 COMPETENCY-BASED SURGICAL EDUCATION ................................................... 12

1.3 ASSESSMENT OF SURGICAL TECHNICAL SKILL ...................................................... 15

1.3.1 VALIDITY AND PSYCHOMETRIC PROPERTIES OF ASSESSMENTS .................... 15

1.3.2 ASSESSMENT IN RESEARCH- OBSERVATIONAL TOOLS ................................... 21

1.3.2.1 Global Rating Scales ......................................................................................... 21

1.3.2.2 Task Specific Checklists .................................................................................. 24

1.3.2.3 Procedure-specific Global Rating Scales ......................................................... 25
1.3.2.4 Error Analysis 29
1.3.2.4.1 Background 29
1.3.2.4.2 Methods 30
1.3.2.4.3 Results 32
1.3.2.4.4 Discussion 47
1.3.3 IN-TRAINING ASSESSMENT – WORKPLACE-BASED ASSESSMENTS 54
1.3.4 SELF-ASSESSMENT 56
1.4 SURGICAL TRAINING IN THE OPERATING ROOM- INSTRUCTIONAL APPROACHES 59
1.4.1 COMMUNICATION AS AN EDUCATIONAL STRATEGY 59
1.4.2 BRIEFING AND DEBRIEFING 64
1.4.3 FEEDBACK 67
1.4.4 VIDEO-FEEDBACK 71
1.4.5 COACHING AND MENTORING 74
1.4.6 TASK DECONSTRUCTION - GRADED RESPONSIBILITY 78
1.4.7 ERROR FOCUSED TRAINING CURRICULA 80
1.5 LIMITATIONS OF LEARNING IN THE OPERATING ROOM 83
1.5.1 ORGANIZATIONAL FACTORS - IMPACT OF WORK-HOUR RESTRICTIONS 83
1.5.2 SOCIETAL FACTORS - PATIENT SAFETY 85
1.5.3 SURGEONS AS EDUCATORS 88
1.5.4 RESIDENTS AS LEARNERS 91
1.6 SUMMARY 93

2 PROBLEM STATEMENT, RESEARCH OBJECTIVES, AND HYPOTHESES 95

2.1 PROBLEM STATEMENT/RATIONAL 96
2.2 RESEARCH AIMS AND OBJECTIVES 98
2.3 HYPOTHESES 99
2.3.1 NULL HYPOTHESIS (H₀) 99
2.3.2 ALTERNATE HYPOTHESIS (Hₐ) 99

3 ERROR RATING TOOL TO IDENTIFY AND ANALYZE TECHNICAL ERRORS AND EVENTS IN LAPAROSCOPIC SURGERY 100
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Abstract</td>
<td>101</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>102</td>
</tr>
<tr>
<td>3.3 Methods</td>
<td>103</td>
</tr>
<tr>
<td>3.4 Results</td>
<td>107</td>
</tr>
<tr>
<td>3.5 Discussion</td>
<td>113</td>
</tr>
<tr>
<td><strong>4 International Consensus on Safe Techniques and Error Definitions in Laparoscopic Surgery</strong></td>
<td>116</td>
</tr>
<tr>
<td>4.1 Abstract</td>
<td>117</td>
</tr>
<tr>
<td>4.2 Introduction</td>
<td>118</td>
</tr>
<tr>
<td>4.3 Methods</td>
<td>119</td>
</tr>
<tr>
<td>4.4 Results</td>
<td>121</td>
</tr>
<tr>
<td>4.5 Discussion</td>
<td>132</td>
</tr>
<tr>
<td>4.6 Conclusion</td>
<td>136</td>
</tr>
<tr>
<td><strong>5 Characterizing &quot;Near Miss&quot; Events in Complex Laparoscopic Surgery through Video-Analysis</strong></td>
<td>137</td>
</tr>
<tr>
<td>5.1 Abstract</td>
<td>138</td>
</tr>
<tr>
<td>5.2 Introduction</td>
<td>139</td>
</tr>
<tr>
<td>5.3 Methods</td>
<td>140</td>
</tr>
<tr>
<td>5.4 Results</td>
<td>143</td>
</tr>
<tr>
<td>5.5 Discussion</td>
<td>146</td>
</tr>
<tr>
<td>5.6 Conclusion</td>
<td>150</td>
</tr>
<tr>
<td><strong>6 Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial</strong></td>
<td>151</td>
</tr>
<tr>
<td>6.1 Abstract</td>
<td>152</td>
</tr>
<tr>
<td>6.2 Introduction</td>
<td>153</td>
</tr>
<tr>
<td>6.3 Methods</td>
<td>154</td>
</tr>
<tr>
<td>6.4 Results</td>
<td>159</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>6.5</td>
<td>167</td>
</tr>
<tr>
<td>6.6</td>
<td>173</td>
</tr>
<tr>
<td>7</td>
<td>174</td>
</tr>
<tr>
<td>7.1</td>
<td>175</td>
</tr>
<tr>
<td>7.2</td>
<td>179</td>
</tr>
<tr>
<td>7.3</td>
<td>182</td>
</tr>
<tr>
<td>7.4</td>
<td>186</td>
</tr>
<tr>
<td>7.5</td>
<td>189</td>
</tr>
<tr>
<td>7.6</td>
<td>191</td>
</tr>
<tr>
<td>8</td>
<td>193</td>
</tr>
<tr>
<td>8.1</td>
<td>194</td>
</tr>
<tr>
<td>8.2</td>
<td>196</td>
</tr>
<tr>
<td>9</td>
<td>199</td>
</tr>
<tr>
<td>9.1</td>
<td>200</td>
</tr>
<tr>
<td>9.2</td>
<td>202</td>
</tr>
<tr>
<td>9.3</td>
<td>203</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>205</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>237</td>
</tr>
<tr>
<td>COPYRIGHT ACKNOWLEDGEMENTS</td>
<td>238</td>
</tr>
</tbody>
</table>
List of Tables

Table 1: Objective Structured Assessment of Technical Skills (OSATS) global rating scale ..... 22

Table 2: Overview of selected articles................................................................. 36

Table 3: Studies conducted to explore the rate and nature of surgical technical error .......... 40

Table 4: Studies using error descriptions as a surrogate for surgical skill in training/ educational setting.................................................................................................................. 44

Table 5: Operative steps of laparoscopic Roux-en-Y gastric bypass ................................ 106

Table 6: Number of errors and events in 25 procedures............................................ 108

Table 7: Error frequency in relation to surgeon’s skill level as determined by the Objective Structured Assessment of Technical Skill global rating scale .................................................. 111

Table 8: Organizations identified and screened for expert inclusion .......................... 122

Table 9: Included expert panel.................................................................................. 123

Table 10: Level of consensus for each survey round and final item selection ............. 127

Table 11: Consensus list included items..................................................................... 130

Table 12: Consensus list excluded items .................................................................. 131

Table 13: Examples of rectification measures ......................................................... 143

Table 14: Overview of Operative steps of LRYGB................................................. 145

Table 15: Demographic details.................................................................................. 160

Table 16: Correlation between baseline demographic data and baseline skill parameters ...... 161

Table 17: Concurrent surgical exposure during a two month rotation ...................... 163
Table 18: OSATS, BOSATS scores, and error counts between and within groups .................. 165
List of Figures

Figure 1: Framework of learning as described by Teunissen et al. (2007)................................. 6

Figure 2: Search and exclusion algorithm in the PRISMA format (Liberati, et al., 2009).......... 33

Figure 3: Examples of potential error-event outcome relationship ........................................ 51

Figure 4: Error-event-outcome hierarchy, with examples. ...................................................... 104

Figure 5: Errors per procedure noted by two observers....................................................... 109

Figure 6: Distribution of error modes for each observer. ...................................................... 110

Figure 7: CONSORT 2010 Flow Diagram (Moher, et al., 2001) ............................................. 162
List of Appendices

Appendix 1: Generic Error Rating Tool (GERT) Checklist (Bonrath, Zevin, et al., 2013)……237
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Abstract Conceptualization</td>
</tr>
<tr>
<td>ACGME</td>
<td>Accreditation Council for Graduate Medical Education</td>
</tr>
<tr>
<td>ACS NSQIP</td>
<td>American College of Surgeons National Surgical Quality Improvement Program</td>
</tr>
<tr>
<td>AERA</td>
<td>American Educational Research Association</td>
</tr>
<tr>
<td>APA</td>
<td>American Psychological Association</td>
</tr>
<tr>
<td>AE</td>
<td>Active Experimentation</td>
</tr>
<tr>
<td>BL</td>
<td>Baseline</td>
</tr>
<tr>
<td>BOSATS</td>
<td>Bariatric Objective Structured Assessment of Technical Skill</td>
</tr>
<tr>
<td>CBD</td>
<td>Common Bile Duct</td>
</tr>
<tr>
<td>CE</td>
<td>Concrete Experience</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CPSO</td>
<td>College of Physicians and Surgeons of Ontario</td>
</tr>
<tr>
<td>CSC</td>
<td>Comprehensive Surgical Coaching</td>
</tr>
<tr>
<td>CT</td>
<td>Conventional Training</td>
</tr>
<tr>
<td>DOPS</td>
<td>Surgical Direct Observation of Procedural Skills</td>
</tr>
<tr>
<td>EAES</td>
<td>European Association for Endoscopic Surgery</td>
</tr>
<tr>
<td>EEM</td>
<td>External Error Mode</td>
</tr>
<tr>
<td>GERT</td>
<td>Generic Error Rating Tool</td>
</tr>
<tr>
<td>GJ</td>
<td>Gastrojejunostomy</td>
</tr>
<tr>
<td>GOALS</td>
<td>Global Operative Assessment of Laparoscopic Skills</td>
</tr>
<tr>
<td>i.q.r</td>
<td>Interquartile Range</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass Correlation Coefficient</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ISCP</td>
<td>Intercollegiate Surgical Curriculum Programme</td>
</tr>
<tr>
<td>JJ</td>
<td>Jejunojejunostomy</td>
</tr>
<tr>
<td>LRYGB</td>
<td>Laparoscopic Roux-en Y Gastric Bypass</td>
</tr>
<tr>
<td>MB</td>
<td>Measuring Bowel in LRYGB</td>
</tr>
<tr>
<td>MIS</td>
<td>Minimally Invasive Surgery</td>
</tr>
<tr>
<td>NCME</td>
<td>National Council on Measurement in Education</td>
</tr>
<tr>
<td>OCHRA</td>
<td>Observational Clinical Human Reliability Analysis</td>
</tr>
<tr>
<td>OR</td>
<td>Operating Room</td>
</tr>
<tr>
<td>OSATS</td>
<td>Objective Structured Assessment of Technical Skill</td>
</tr>
<tr>
<td>PBA</td>
<td>Procedure-based Assessment</td>
</tr>
<tr>
<td>PGY</td>
<td>Postgraduate Year</td>
</tr>
<tr>
<td>pouch</td>
<td>Creation of Gastric Pouch in LRYGB</td>
</tr>
<tr>
<td>PRISMA</td>
<td>Preferred Reporting Items for Systematic Reviews and Meta-analyses</td>
</tr>
<tr>
<td>PT</td>
<td>Post-Training</td>
</tr>
<tr>
<td>r</td>
<td>Pearson’s r</td>
</tr>
<tr>
<td>RACS</td>
<td>Royal Australasian College of Surgeons</td>
</tr>
<tr>
<td>RO</td>
<td>Reflective Observation</td>
</tr>
<tr>
<td>rs</td>
<td>Spearman’s Rho</td>
</tr>
<tr>
<td>SAGES</td>
<td>Society of American Gastrointestinal and Endoscopic Surgeons</td>
</tr>
<tr>
<td>U.K.</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>U.S.</td>
<td>United States</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual Reality</td>
</tr>
</tbody>
</table>
THESIS OVERVIEW

This thesis follows a paper format.

Chapter 1: Literature Review

In this chapter, I will describe the context in which surgical learning takes place. As this thesis focuses on technical skill development, aspects related to acquiring surgical skill in residency training will be detailed. First, I will explain the theoretical principles of experiential learning that contribute significantly to skill acquisition in a workplace-based learning environment. Then, I will describe three models of postgraduate surgical training, to highlight the background of surgical training, and to demonstrate that modern programs still share numerous commonalities with older models. As assessments form the foundation of modern competency-based training and serve as a source of feedback for reflective observation in experiential learning, I will continue by describing observational assessment methods used in research and in-training settings. Validity of evaluations depends substantially on constructs and principles underlying the scales used as well as on the psychometric properties of the scales applied. I will thus, detail the two relevant frameworks of validity relevant to surgical observational assessments. The focus of this work is on surgical training in the operating room (OR); therefore, current instructional approaches will be reviewed and discussed. Lastly, I will highlight some of the barriers to learning in the OR.

The contents of section 1.3.2.4 have been published in "Defining Technical Errors in Laparoscopic Surgery: A Systematic Review" by Bonrath EM, Dedy NJ, Zevin B, Grantcharov TP, Surgical Endoscopy August 2013; 27(8):2678-2691.
Chapter 2: Problem Statement, Research Objectives, and Hypotheses

In this chapter, I will describe the knowledge gap addressed in the current thesis. I will also state the main research aim, objectives and formulate the hypothesis that guided this work.

Chapter 3: Error Rating Tool to Identify and Analyze Technical errors and Events in Laparoscopic Surgery

The contents of this chapter have been published in “Error rating tool to identify and analyse technical errors and events in laparoscopic surgery” Bonrath EM, Zevin B, Dedy NJ, and Grantcharov TP, British Journal of Surgery, 2013;100(8):1080-8. The chapter describes the design of an error assessment framework, as well as the validity evidence gathered through applying the framework to video recordings of laparoscopic Roux-en-Y gastric bypass (LRYGB) procedures.

Chapter 4: International Consensus on Safe techniques and Error Definitions in Laparoscopic Surgery

The contents of this chapter have been published in “International consensus on safe techniques and error definitions in laparoscopic surgery” Bonrath EM, Dedy NJ, Zevin B, and Grantcharov TP, Surgical Endoscopy 2014; 28(5):1535-1544. The chapter describes a Delphi consensus process conducted with expert laparoscopic surgeons and educators to establish a reference list of error definitions and examples pertinent to surgical training.
Chapter 5: Characterizing “Near Miss” Events in Complex Laparoscopic Surgery through Video-Analysis

The contents of the chapter have been published in “Characterising “Near Miss” Events in Complex Laparoscopic Surgery through Video-Analysis”, Bonrath EM, Gordon LE, and Grantcharov TP, BMJ Quality and Safety, Online First 6 May 2015, doi:10.1136/bmjqs-2014-003816. The chapter describes intraoperative events identified through video-review of successful routine procedures. The study aimed to determine the presumed underlying error mechanism causative of the observed events. The reported study takes an explicit look at injury mechanisms to create an evidence base for the teaching of technical skills founded on common examples of error- injury patterns. The analysis informs the educator of the specific learning environment and factors that should be addressed during intraoperative teaching.

Chapter 6: Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial

The contents of the chapter have been published in “Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial”, Bonrath EM, Dedy NJ, Gordon LE, and Grantcharov TP, Annals of Surgery, 2015. In this chapter, the randomized controlled trial that was conducted to address the primary research aim of this thesis is described. The effectiveness of a surgical coaching intervention, which includes the instructional approaches of debriefing, feedback, and behavior modeling, is compared to the effectiveness of conventional residency training.

Chapter 7: General Discussion

In this chapter, I will summarize and discuss the general findings of the studies reported in the prior chapters. Main discussion points include the importance of comprehensive objective assessment in training, the need for evidence-based and content standardized intraoperative
teaching, and general considerations for technical skills training in residency building on coaching frameworks. Furthermore, the relevance of self-assessment in surgical training and self-directed learning will be discussed.

Chapter 8: Limitations

In this chapter, I will address the most pertinent limitations to the research reported in the thesis. The main limitations are the inability to correlate our findings to patient outcome data through the nature of the study designs, and limitations regarding generalizability of findings to other educational environments.

Chapter 9: Future Directions

In this final chapter, I will summarize the directions that future research could take, building on the work reported in this thesis. Each study described, can lead on to future research questions, some of which may address the limitations detailed in the previous chapter.
In order to understand surgical postgraduate training, it is necessary to review the context in which learning takes place. The learning environment, as well as the structures surrounding how and when teaching is delivered, need to be considered when aiming to devise new training approaches for surgical postgraduate education. Surgical trainees underlie the pressures of dual, occasionally conflicting, roles. These roles include the need to provide high quality medical care during their everyday professional duties as well as constraints due to the need to personally grow as a professional, expanding their knowledge and skills through structured “curricular”, and unstructured “on the job” learning. Subsequently in this chapter, I have detailed topics pertinent to understanding this complex learning environment. The focus of this thesis will be on technical skills development in laparoscopic surgery and, for this reason, only technical aspects will be the subject of this overview. The contents of this chapter have been derived from a thorough review of the literature using entries registered in MEDLINE(R) (In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R)), Evidence-Based Medicine (EBM) Reviews (OVID) including Cochrane Database of Systematic Reviews, Database of Abstracts of Reviews of Effects, Cochrane Central Register of Controlled Trials, Cochrane Methodology Register, Health Technology Assessment Database, NHS Economic Evaluation Database, and Embase. Additional sources were identified through cross-references to relevant articles, as well as through review of grey literature (websites) and book texts.
1.1 Experiential Learning Theory

The understanding that individuals can learn from the experiences they make, is embedded in cultures utilizing apprenticeships to teach crafts and trades to the next generation at least since the Middle Ages (Hansen, 2004). In the experiential learning model, the workplace represents the core learning environment, which aims to reinforce formal education (Kolb, 1984). The opinion, that subjective experiences may be valuable to the learning process, contrasted the view on traditional transmission education that valued only the objective and factual. Pioneers, such as John Dewey, Kurt Lewin and Jean Piaget, explored the influence of subjective individual experiences on the learning process and thus created the broad knowledge base on which most of contemporary experiential learning theory, as popularized by David Kolb, is currently based (Kolb, 1984). Modern applications of experiential learning theory include competency-based education, lifelong learning and professional development, internships, and simulations (Kolb, 1984).

Under the presumption that learning represents a continuous process, David Kolb defined four distinct learning modes, described as “abilities” that represent repeating elements encountered during experiential learning. These are:

“concrete experience abilities (CE), reflective observation abilities (RO), abstract conceptualization abilities (AC), and active experimentation (AE) abilities.” (Kolb, 1984)[pg 30].

During concrete experience the learner performs or experiences a new task or situation, it is this experience from which the learning will be derived. Reflective observation describes processes where the learner re-explores the prior experience; this may include self-reflection or discussion with a facilitator or peers. Abstract conceptualization aims to integrate the new experience into pre-existing knowledge, and lastly the learner seeks to apply the newly acquired knowledge
during active experimentation. Learning is a continuous cyclic process, and thus these learning modes are generally not exhibited all at the same time, yet they must not always follow the same order (Kolb, 1984). Learners may need abstract, theoretical skills such as conceptualization and reflective abilities in one setting, and then shift to using the active abilities of experimentation or experiencing in another. David Kolb (1984) compared his interpretation of the learning cycle with similar constructs targeting problem-solving, decision-making and scientific inquiry and highlighted that most processes follow similar paths, although the descriptors of each stage may vary. Kolb's interpretation of the experiential learning process is heavily influenced by John Dewey's model, which identifies that an experience in itself does not constitute learning, but rather that it is the reflective process and the potential consequences of an action that represent the learning potential (Dewey, 1916). Subsequently, modern applications of experiential learning build on the fundamental understanding that an experience initiates learning, modified by subjective interpretation and subsequent generalization of concepts for future experiences. Workplace-based training, as in surgical education, essentially will to some degree always draw on concepts of experiential learning throughout the learning process.

Although concepts of experiential learning are well established in adult education, the theory has also attracted criticism. The criticism voiced is due to the perceived lack of structure and guidance in educational approaches such as discovery learning or problem-based learning, that are deemed to be based on experiential learning theory (Kirschner, Sweller, & Clark, 2006). The assumption that the act of problem-solving, as an educational activity, draws heavily on the learner's working memory, which is presumed to inhibit learning, has been explored in several settings investigating the effectiveness of "exploratory" or "discovery" learning exercises (Klahr & Nigam, 2004; Moreno, 2004; Tuovinen & Sweller, 1999). These studies though involved a "no guidance" treatment group termed “discovery-learning” that was compared to a group receiving feedback or guidance. Both groups participated in active hands-on learning exercises, thus learned from an experience, but in the “exploratory learning groups” reflection and abstract conceptualization was left to the learner without any guidance by a facilitator (Klahr & Nigam, 2004; Moreno, 2004; Tuovinen & Sweller, 1999). Furthermore, the detrimental effect of “no guidance” on learning was most pronounced in individuals that had no prior experience in the skill to be learnt (complete novices), and was not measurable in individuals that had a knowledge base on which to build the new experiences (Tuovinen & Sweller, 1999). As suggested in the
experiential learning theory, the learner aims to conceptualize new content by embedding the knowledge in already existing frameworks, for this reason without any prior experience self-guided learning may be difficult. Furthermore, in the study by Roxana Moreno (2004), the variable that was manipulated was the form of feedback offered to learners. The group receiving explanatory feedback outperformed the individuals that only received corrective feedback without explanations (“correct” or “incorrect”) (Moreno, 2004). The relevance of feedback has been identified as central to learning and will, therefore, be discussed in detail further on in this chapter. Therefore, although Kirschner and colleagues make a compelling argument against constructivist experiential learning models, in fact the criticism needs to be addressed towards a lack of guidance and structure especially in novice learners. In the same vein, Mayer (2004) summarized seminal research on discovery learning methods from three decades, concluding that in numerous differing settings pure discovery without guidance was inferior to guided discovery, suggesting that in constructivist learning, teaching approaches that include structure and guidance are of high importance to ensure learning can actually occur. Therefore, it does not appear reasonable to equate experiential learning simply to unguided “learning on the job”. In a recent study, unguided “on the job learning” was compared to learning in a structured curriculum, and was subsequently shown to be inferior (Bradley, Webb, Schmitz, Chipman, & Brasel, 2010). As discussed previously, exposure to daily experiences without the steps of critical reflection, applying judgment and conceptualization of the learning content do not constitute experiential learning. Therefore, although residency training is a good environment to employ techniques of experiential learning to build knowledge, experiences that remain without reflection will not necessarily contribute to the learning process.

Thus, to best benefit from practical experiences, surgical training should include guidance throughout the learning process, with trainers employing instructional methodologies such as facilitated self-reflection in debriefing, and peer or trainer feedback.

A study conducted in 2007 in the Netherlands, thoroughly examined processes by which obstetrics and gynaecology residents learned within a clinical environment (Teunissen et al., 2007). Following a grounded theory approach, a framework of learning was defined that was deemed applicable to resident's learning in clinical rotations (Figure 1). Focus group discussions with residents revealed that learning started through actively participating in work-related duties. These experiences were subsequently interpreted as leading to the creation of “personal
These personal experiences represented the basis for knowledge construction, resulting in the creation of a “personal knowledge” base (Teunissen, et al., 2007). The process of self-reflection was identified to further the personal knowledge, which could be initiated through external sources such as supervisor evaluations or internal sources such as pre-existing knowledge. In addition, internal and external factors were postulated to influence learning at all stages of the process (arrows depicted in Figure 1). For example, depending on prior knowledge (arrow A₃ in Figure 1), structure of training environment (arrow E₂ in Figure 1), and nature of interactions with peers and supervisors interpretation of activities (arrows A₂, B₂, C₂ in Figure 1), construction of meanings and self-reflection mechanisms could be modified (Teunissen, et al., 2007). The framework of learning developed by Teunissen et al. (2007) bears many similarities to the processes attributed to the experiential learning cycle. The core similarities between the two learning modes are the central role of the “activity”, the fact that interpretation of an activity leads to construction of new knowledge, and that reflection on activities and knowledge furthers the learning process. The work by Teunissen et al. (2007) provided evidence to support the hypothesis that residents do, in fact, learn from clinical activities during routine work. Since numerous external factors were deemed to influence the learning processes including interactions with supervisors and educators, the authors emphasized that faculty need to be aware of the impact their actions and opinions may have on resident learning (Teunissen, et al., 2007).
Figure 1: Framework of learning as described by Teunissen et al. (2007)

Reprinted with kind permission from John Wiley and Sons, Inc.


The learning processes derived from active participation in work activities are depicted as internal steps (solid lines) with dotted lines representing the internal regulations and modifiers. External factors could influence the learning processes at all stages and are depicted as labeled solid arrows (Xa).
1.2 Models of Postgraduate Surgical Education

In this dissertation, I will use the terms “surgical trainees” and “surgical residents” interchangeably. Both these terms refer to the individuals learning the profession of a surgeon. Since the processes surrounding how an individual acquires the skills necessary to practice the trade of a surgeon vary between cultures, surgical education cannot be generalized and simplified to one “standard” training program. Singh et al. (2014) examined the differences between just a few select training programs in a national comparison, revealing numerous commonalities, but also some significant differences. The training programs reviewed all combined workplace-based learning following principles of experiential learning with structured curricular components taught outside the workplace environment (Singh, Aggarwal, & Darzi, 2014). The clearest difference between these training programs was the length of training which varied from five to ten years (Singh, Aggarwal, & Darzi, 2014). Although the programs examined followed modern conceptions of competency-based training, which included descriptions of curriculum content, training requirements and competency objectives, elements from older training structures were also still prevalent (Singh, Aggarwal, & Darzi, 2014). For example, to some degree the programs were still found to adhere to volume- and time criteria to determine progression, despite all programs emphasizing the role of competence assessments. Thus, as educational concepts have evolved over time, even modern programs retain evidence of early foundations including concepts of apprenticeship style training. Subsequently, in the following sections, I will discuss the dominant training paradigms that represent the basis for modern surgical training.
1.2.1 Apprenticeship Model

The apprenticeship represents a core instructional paradigm where the novice is introduced to the realm of the expert and by actively participating in this environment gradually becomes the expert himself (Farnham-Diggory, 1994). Apprenticeships evolved from the medieval custom of sending young children away from home to host families in order to learn manners and to be instructed in skills required for daily life. From this custom, through the sixteenth and seventeenth century, apprenticeships in England became a common educational method to teach skills required in technical trades (Lane, 1996). Apprentices were young, and bound to a “master” for a predefined term of usually seven years. In surgery, which originated from a trade profession, apprenticeships were also common (Lane, 1996). Although surgical apprentices were generally recruited from families with higher social standing than those recruited for other apprenticeships. In addition, terms and conditions of surgeon apprenticeships were better than those for other trade professions (Lane, 1996). In the eighteenth century, the status of the surgical profession was elevated by medical schools to that similar of “physicians”, with the surgical trade being recognized as an important part of the medical professions (Lane, 1996). Apprenticeships though were still continued. An article from The British Medical Journal published in 1891 detailed that two systems of training coexisted and in part conflicted in medicine. Those individuals that pursued a less academic path in traditional apprenticeships were looked down upon by the “scientific students.” On the other hand, apprentices believed that the “scholars” would fail in practice (BMJ, 1891). As lectures in medical training became compulsory the pure apprentice role was exchanged against a combination of practical training through affiliation with a “master” whilst still following the scholarly duties in medical training (BMJ, 1891).

Today apprenticeship still represents a core aspect of surgical postgraduate education. The apprentice learns over a prolonged period from observing the “master.” Following principles of graded responsibility, the apprentice is afforded opportunities to at first complete steps of a procedure under supervision and subsequently with growing experience goes on to complete entire procedures under supervision (Traynor, 2011). The level of oversight deemed necessary by the supervising “master” is gradually reduced until the apprentice is considered competent for
independent practice. This process requires a close bond between “master” and “apprentice” and requires extensive opportunities for the apprentice to observe procedures (Traynor, 2011). Although many aspects of traditional apprenticeships have been replaced by curricular residency programs, the implementation of work-hour restrictions in the United States (U.S.) has led educators to revisit components of the apprenticeship model (Darosa, Bell, & Dunnington, 2003; Heitmiller, Gupta, & You, 2008; Schneider, Coyle, Ryan, Bell, & DaRosa, 2007). In a study by Schneider et al. (2007) the apprenticeship rotations developed within a new work-hour compliant residency model were rated more positively by residents than other rotation models. Specifically, aspects of teaching quality and continuity of patient care, were evaluated more favorably for apprenticeship rotations compared to team rotations (Schneider, et al., 2007). Similarly, a “junior apprentice resident” rotation was introduced at the Union Memorial Hospital in Baltimore to expose junior trainees to general and colorectal surgery in a private practice setting (Heitmiller, et al., 2008). This one-month rotation ensured that junior trainees experienced continuity of care working alongside one high-volume surgeon and were afforded an opportunity to participate actively in operative cases without competing for these with more senior residents. The apprenticeship model was well received, resulted in higher case volumes and was found to comply easily with the 80-hour work directive (Heitmiller, et al., 2008). Both these programs highlight the benefits of apprentice elements in modern residency. Specifically, the stronger trainee-trainer bond and supervision were identified as benefits that resulted in a perceived higher quality of teaching and continuity of care. Therefore, it does not seem surprising that a recent survey of vascular surgery residents also revealed that the residents believed that they learned most effectively when being in an apprenticeship-style learning environment receiving direct instruction from a designated supervisor (Dalsing et al., 2012). One caveat mentioned by Schneider et al. (2007) though, was that apprenticeship models required significant faculty involvement. Also, evaluation of the individual’s teaching skills and practice profile are necessary before a particular faculty member should be matched with an apprenticeship rotation (Schneider, et al., 2007).
1.2.2 Early Residencies and the Role of Professor William Halsted and Professor Edward Churchill

In the late 1800s up to the early twentieth century, surgery as a field of medical specialization gradually evolved (Rutkow, 2013). At this time, practitioners in the U.S. had little opportunities to gather hands-on experience in the surgical craft following graduation from medical school (Rutkow, 2013). There was no structure in place to guide their surgical training or guarantee sufficient surgical exposure (Rutkow, 2013). Early institutions within the U.S. that aimed at enabling a formalized hands-on exposure to surgical conditions under the mentorship of renowned surgeons included the New York Polyclinic and the New York Post-Graduate School (Rutkow, 2013). Although these institutions offered dedicated exposure and training to enrolled individuals, the duration of the training was brief. At the same time, numerous prospective surgeons left the U.S. to seek education abroad, especially in Germany and Austria-Hungary (Rutkow, 2013). In this era, Dr. William Halsted, a young physician also chose to travel to Europe to gain experience in surgery. Following this journey, Dr. Halsted was heavily influenced by the German training system after which he modeled his conception of post-graduate surgical training (Kerr & O'Leary, 1999). The training approach popularized by Professor Halsted in 1889 at John's Hopkins Hospital (Baltimore) represented formal training based on an apprenticeship model that valued surgical excellence. Through this training model, Professor Halsted strived to educate not only highly qualified medical practitioners but to create scholars and educators (Cameron, 1997). This residency concept consisted of a one-year internship followed by eight higher education years. Progression from one level to the next was determined by Professor Halstead which resulted in far longer training periods than initially intended (Grillo, 2004; O'Sullivan, 2008). Since promotion was determined by Professor Halsted himself, the model developed into a pyramid form with a wide base of junior, and only a few select, senior trainees. Only these senior trainees were fully entrusted with the care of Professor Halsted’s patients (O'Sullivan, 2008; Rutkow, 2013).

Although the Halstedian conception of residency aimed to emphasize scholarly activity as a central part of becoming a surgeon, many aspects of apprenticeships remained. Specifically the fact that trainees were educated by one single experienced surgeon and thus acquired the knowledge, skills and attitudes conveyed by that individual. This single-surgeon centered
education system later attracted criticism, with educators such as Professor Edward Churchill advocating education by a group of people acting as teachers (Grillo, 2004). Professor Churchill subsequently led the way towards the structure of modern “rectangular” residencies, establishing a training concept that aimed to ensure that all selected trainees completed designated training for a minimum of five years (Grillo, 2004). The educational principles underlying Professor Churchill's conception of surgical training included avoiding intern exploitation, competency oriented education with flexibility within the curriculum to ensure proficiency development, and principles of graded responsibility (Grillo, 2004). In addition, this model focused on basic science as an integrated component of surgical education. This was in contrast to the Halstedian model that mandated that prospective surgeons partake in original research throughout their training (Grillo, 2004). Also surgical education was to be structured by means of a graduate education advisory board, and the role of a surgeon educator in staff appointments was emphasized (Grillo, 2004). Beyond the five-year structure, Professor Churchill also included an additional two-year extra training term for those individuals aiming to become surgical professors (Grillo, 2004). The new training program was initiated in 1946 and has continued to serve as the dominating surgical training model in the U.S. since (Grillo, 2004).

Following Halsted's and Churchill's reformations of surgical education in the past century, modern residency continues to maintain aspects of the apprenticeship model, such as the principle of graded responsibility and learning from an experienced individual, especially during interactions in the OR, but it also involves significantly more structure and curricular content. Residency is currently undergoing another paradigm shift away from a time-based system towards an outcomes-orientated competency-based system.
1.2.3 Competency-Based Surgical Education

Modern competency-based education and training has its origins in teacher education where demands of higher levels of accountability regarding the quality of training of elementary school teachers resulted in a re-examination of curricula and certification processes (Burke, 1989). Examples of core aspects of competency-based education include, that specific competencies reflecting the roles or characteristics of a profession need to be clarified and made public (Burke, 1989). Further, those competencies should be predictive of professional effectiveness (Burke, 1989). With regards to assessment, the measures chosen to determine competency should be realistic, valid, and based on predefined standards set for each competency with competency statements allowing for criterion referencing (Burke, 1989).

In the health professions, competency-based education has recently re-emerged displacing many of the older training paradigms described in previous sections. Although the concept was already embedded in Professor Churchill’s vision of surgical residency, competency-based education at the time did not find implementation. A systematic review of the literature focusing on articles including definitions of competency-based education revealed a sharp increase in the number of articles over the past decade (Frank et al., 2010). Since there has been much debate about what the term “competency-based education” in medicine means, Frank et al. (2010), reviewed articles detailing definitions of competency-based medical education and subsequently proposed a new definition applicable to medical education. The proposed definition was:

“competency-based education (CBE) is an approach to preparing physicians for practice that is fundamentally orientated to graduate outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It de-emphasizes time-based training and promises greater accountability, flexibility, and learner-centredness” (Frank, et al., 2010) [pg 636].
Predefined relevant competencies thus represent the core of any competency-based training program. In surgical education, the professional competencies have been defined by the respective governing bodies such as the Royal College of Physicians and Surgeons of Canada (Frank JR, 2014), The Royal Australasian College of Surgeons (RACS) (Royal Australasian College of Surgeons, 2014), and the Accreditation Council for Graduate Medical Education (ACGME) in the U.S. (ACGME & American Board of Surgery, 2014). The competencies that have been defined are general attributes of a surgeon, and although several competency frameworks have been broken down into more detailed sub-competencies, classification of everyday activities into these subgroups may still remain challenging (ten Cate & Scheele, 2007). As advocated by ten Cate and Scheele (2007) educators should thus aim to focus competency assessments on selected activities that reflect the professional work of the surgeon. These activities provide evidence of competence within the overarching general competence attribute defined by the regulatory bodies. Nevertheless, the authors also caution that

“the competent professional is clearly more than the sum of many detailed operational competencies” (ten Cate & Scheele, 2007)[pg 543].

Another central element of competency-based training is the need to assess and objectively judge competencies in order to determine whether they have been sufficiently mastered. Determining competence though may be challenging, since not only do explicit benchmarks need defining, but furthermore reliable assessment methodologies are required to determine whether these benchmarks have been met. Several essential conditions must be fulfilled to ensure that evaluations in competency-based training are fair and representative. These include the need for frequent and continuous assessment processes that are conducted in the environment in which the competency is to be exhibited (workplace-based assessments), appropriate criterion-referencing of the measurements, ensuring that standardized tools are used, and deriving assessments from multiple sources including the trainee him/herself (Holmboe, Sherbino, Long, Swing, & Frank, 2010). Furthermore, assessments in competency-based training not only serve the purpose of determining when an individual can progress within the program, but also provide a means of delivering feedback to the learner as to guide future learning. In addition, in a system, that is solely outcomes-orientated, timely identification of weaknesses to facilitate early remediation is also of high importance. Thus, both summative and formative feedback play a
central role in all competency-based programs (Holmboe, et al., 2010). Since assessments play a pivotal role in competency-based education, it is also the lack of genuine, meaningful assessment that poses the greatest threat to the educational experience in competency-based programs. The need for direct faculty observation in competency-based programs remains problematic, and a lack thereof has been perceived as one of the major barriers to successful implementation of competency-based programs (Holmboe, 2014).
1.3 Assessment of Surgical Technical Skill

As detailed in the previous section, the quintessence of competency-based training and a core aspect of experiential learning alike is the ability to assess performance. Although surgical competence encompasses a wide set of skills and abilities, the focus of the present work is on surgical technical performance and skill. Thus, in the next sections I will discuss assessment methodologies designed to evaluate surgical technical performance in an OR environment. Since direct observation of performance plays a central role in modern competency-based programs, I will focus the following discussion on observational tools. A number of device-based performance measures such as motion tracking and grasp force measurements have also been developed, but as they require specialized equipment and have not found widespread implementation, they will not be subject of this overview. Workplace-based assessments represent practical applications of observational assessments in competency-based training and will thus be discussed separately from assessment methodologies predominantly used in a research setting. In addition, as an assessment is only of value if the test scores arising from the assessment are interpreted correctly, I will also describe the main principles of determining test-score validity within the following sections of this chapter.

1.3.1 Validity and Psychometric Properties of Assessments

Assessment instruments in surgical education are used to quantify a specific surgical attribute, attitude, behavior, or skill. In the most cases, the tools measure a particular “construct”, which unlike other measurements in science such as temperature, weight, heart rate etc., represent notions or ideological frameworks that are hard to quantify (Downing, 2003). Constructs may thus only be measurable through indirect inferences (Downing, 2003). Subsequently, evidence must be gathered that supports the correctness of the assumption that the indirect measurement represents the construct of interest to a degree that the test results can be trusted. The measurements can be used in a formative fashion, to determine learning needs and serve as a
source for feedback, or for summative purposes to determine progression from one training stage to the next. In both settings, it is important to ensure that the measurements made are accurate (reliable), and that they measure what they were designed to measure (valid). Naturally, the more consequential an assessment becomes, the stronger the evidence to support assumptions of reliability and validity will need to be. Subsequently, end-users of any measurement tool will need to be convinced, that for the purposes and setting in which they aim to apply the tool, enough evidence supports the interpretation of the result. “Validity studies” aim to provide the evidence required to ensure test-score interpretations are correct and meaningful. As reiterated by Samuel Messick in his seminal article advocating a unified concept of validity in 1995, it is important to note that

“validity is not a property of the test or assessment as such, but rather of the meaning of the test scores.” (Messick, 1995b)[pg 741].

The ideologies and framework surrounding validity evidence have undergone a shift with two distinct approaches coexisting in the medical literature. The American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME), advocate a revised approach to validity assessment of new tests, using a unified approach (AERA, APA, & NCME, 1999). Although this framework represents the standard for test and tool design, the framework has not found wide adoption in surgical education (Ghaderi et al., 2014; Korndorffer, Kasten, & Downing, 2010). Recent reviews in surgery have extensively explored the properties of global rating scales and task-specific checklists (K. Ahmed, Miskovic, Darzi, Athanasiou, & Hanna, 2011; Ghaderi, et al., 2014; van Hove, Tuijthof, Verdaasdonk, Stassen, & Dankelman, 2010). These reviews demonstrated that the vast majority of “validity studies” still report according to the old structure. Van Hove et al. (2010) and K. Ahmed et al. (2011) explored validity according to the old framework of face, content, construct validity and reliability, whilst Ghaderi et al. (2014), reviewed assessment tools with the aim to determine the amount of validity evidence gathered for each tool using the new unified framework. These reviews highlight that in the surgical education literature, the vast majority of studies conducted to explore psychometric properties of assessment tools still follow the older framework established in 1966 by the APA (APA, 1966). For this reason, both conceptions of validity will be discussed in this thesis.
The older “traditional” view of validity was established in the 1950's and was anchored in the testing standards published in 1954, updated in 1966 (APA, 1954, 1966). This validity conception was structured around the definition of three distinct validity types: content, criterion-related (either as concurrent or predictive), and construct validities (Messick, 1989). The three validity types were supplemented by a fourth category “reliability” (APA, 1966). The validity type frequently quoted in medical literature of “face validity”, referring to whether a test is viewed to measure what it is supposed to through subjective opinions of test users or third parties, has been discounted as a formal form of validity evidence and is not included in official psychological testing frameworks (Messick, 1989).

Following the older view of validity, content validity refers to the extent that test items reflect the scope of intended content (Zeller & Carmines, 1980). Criterion-related validity defines how a test measure compares to existing established measures relating to the test’s domain and construct. Criterion-related validity can be subdivided into correlations with measures determined at the same time of assessment (concurrent) or against future measures (predictive) (Zeller & Carmines, 1980). For example, if the scores obtained on a new written medical knowledge test were referenced against in-training assessments (as an established assessment measure) which also judge medical knowledge, a positive strong correlation between test scores and in-training assessments would then demonstrate concurrent criterion-related validity for the new written test. Similarly, if the results of this written exam were later correlated to the rate of successful completion of medical training, then the knowledge test would be assumed to show predictive criterion-related validity. Construct validity in contrast to criterion-related validity is less tangible as the measure that is to be explored is theoretical in nature. The interpretation of a construct can lead to the formulation of hypotheses aligning with the assumed construct. In assessing construct validity, measures are explored that are assumed to test the hypotheses derived from the construct of interest (Zeller & Carmines, 1980). For example, a common means of assessing construct validity in assessments of surgical performance is testing the hypothesis that surgeons with a higher level of operative experience (experts) will score higher on scales aimed at measuring surgical skill than surgeons with lower skill levels (novices). Thus, if an expert's test score is significantly higher than that of a novice trainee then the hypothesis would be confirmed and construct validity for the construct that the test measures skill would also be supported.
The view on validity has shifted with regards to the understanding that validity types represent different views on one construct rather than representing distinct entities. In addition, gathering empirical explanatory evidence has been viewed of higher relevance than determining mere predictive ability when considering test validity (Messick, 1989). This shift in interpretation of validity was a gradual process and has taken several decades, with subtle changes noted within the “Standards” of psychological testing throughout edition revisions (Messick, 1989). In 1999, the traditional framework was ultimately revised in collaboration by the APA, AERA, and NCME, to address the new emerging validity concepts (AERA, et al., 1999). Within this revised framework, validity is viewed in a comprehensive manner to include all sources of evidence that support the correct interpretation of test results (Messick, 1989). Since evidence is not absolute, validation is a dynamic process where new evidence supplements existing evidence to aid the interpretation of test results (Messick, 1989). All sources of validity are subsumed under the term “construct validity,” which can further be divided into six aspects for which validity evidence can be gathered (Messick, 1995b). No single category overrides the other so that abundant validity evidence in one category does not preclude the need to collect evidence reflecting the other aspects of construct validity (Messick, 1995b). The six aspects of construct validity defined by Samuel Messick were “content”, “substantive”, “structural”, “generalizability”, “external” and “consequential” aspects (Messick, 1995b). From this framework, the current Standards for Educational and Psychological Testing were developed (AERA, et al., 1999). These “Standards” detail five sources of evidence that can be used to support validity of a measurement tool: “content”, “response process”, “internal structure”, “relationship to other variables” and “consequences” (AERA, et al., 1999).

Content evidence, as in the older framework aims to demonstrate that the items of a scale or test are representative of the construct that the tool is proposed to measure. Sources of content evidence include details on how the test was designed and by whom, the credentials and expertise of the individuals designing the items, clarity of item wording, and representativeness of all domains assumed to reflect the construct of interest (Downing, 2003; Messick, 1995b).

The evaluation of response processes aims to identify errors due to the administration of a test or application of an observational tool (Downing, 2003). For example, although content may be representative of the intended construct, the structure of a scale or wording may illicit an unanticipated response reaction by the rater that is not related to the construct, rather that is a
consequence of the tool design. Therefore, by understanding that different raters may interpret scale items differently, measures can be taken to remove ambiguous terms or descriptive anchors and definitions can be included to guide the rater. Details on rater training and the knowledge gained from reviewing the calibration processes can further be used to inform scale designers of potential threats to validity within the response process. Response processes evidence also includes descriptions of how data is handled to avoid data entry errors, how composite scores were derived, and what rationale was used to establish weighted scores (Downing, 2003).

Internal consistency refers to the psychometric scale properties of a test and can be explored through a variety of statistics (Downing, 2003). Validity evidence from this source is frequently reported in “validation” studies (K. Ahmed, et al., 2011; Ghaderi, et al., 2014; van Hove, et al., 2010). Several different types of reliability can be evaluated to assess the internal structure of a tool. Beckman et al. (2004) summarized the most relevant measures. The internal consistency of a test describes whether the items of a tool pertain to one or more constructs that are to be measured (Beckman, Ghosh, Cook, Erwin, & Mandrekar, 2004). High internal consistency with good inter-item correlation suggests that only one construct is represented within the framework. If an item does not correlate well with the other items within a scale or test this may indicate that the item relates to a different construct. A common statistical measure reported for internal consistency is Cronbach’s alpha. Temporal stability refers to whether a test taken by an individual at two differing time points results in comparable scores, the statistical measure of test-retest reliability are correlation statistics (Beckman, et al., 2004). Test equivalence refers to whether two versions of a test aimed at measuring the same construct result in comparable scores, this form of reliability is also assessed through correlation statistics (Beckman, et al., 2004). Inter-rater agreement appraises whether different observers applying a test or scale produce similar test results, whereas intra-rater agreement, like test-retest reliability, examines the similarity of measurements made by one individual over time. Statistical measures commonly used to express inter-rater reliability are Kappa for categorical data, and Intraclass Correlation Coefficients (ICC) for continuous data (Beckman, et al., 2004). Intra-rater reliability can also be assessed by ICC’s. Last, a comprehensive view of reliability can be achieved by applying generalizability theory, which is a complex statistic that allows the determination of how strongly different factors contribute to the final test score of a measurement tool.
Generalizability studies enable the identification of potential sources of error that need to be accounted for during the use of a tool (Beckman, et al., 2004).

Evidence gathered from the comparison of test measures with other variables include concepts of convergent and divergent validity evidence, predictive validity and criterion-referenced validity evidence. Convergent validity evidence can be obtained through correlation of results obtained by using two tests measuring similar constructs. If both tests measure similar construct, a degree of correlation should be observed. Similarly, divergent validity evidence can be demonstrated by comparing results of two tests that are deemed to measure unrelated constructs, if test scores, in fact, do not correlate (Downing, 2003). This form of validity evidence is commonly used in validation studies in surgical education (Ghaderi, et al., 2014).

Consequential evidence refers to whether and how test results influence decisions at the individual but also societal level. In the context of surgical education, the extent of test score consequence depends on the proposed use of the tool. Tools used with a formative intent will result in the identification of training needs and subsequently errors due to invalidity of test result interpretation may have as a consequence that training content is inappropriately chosen. In contrast invalidity score interpretation of instruments used for summative assessments, such as those used to determine graduation or promotion, may have far reaching consequences. The need for consequential validity evidence though remains a topic of debate. Advocates for consequential validity argue that a test may have societal consequences (positive or negative) and therefore evidence needs to be gathered that ensures the interpretation of a test result is based on the performance rather than an unanticipated threat to validity within the test design, and that potential positive consequences arising from test application are recognized as such (Messick, 1995a, 1995b). Adversaries argue that a test score need not have a consequence, and that decisions pertaining to the use of a given score should not be attributed to evidence to support that the construct was accurately inferred from the test result (Mehrens, 1997).

In summary, it must be noted that elements of the older classification are still represented within the revised unified framework. For example, content aspects of validity still remain a cornerstone of validity assessments, and criterion aspects are summarized under the category of “evidence gathered from comparison to other variables”. In addition, reliability measures are summarized under the category “internal structure” within the unified framework. Although views have
shifted, the fundamental understanding of evidence sources supporting validity remains unchanged. As the majority of surgical education literature still uses the terms associated with the older framework of validity, including the concept of validation studies, these terms will be continued to be used throughout this thesis, supplemented by reference to the unified framework where necessary.

1.3.2 Assessment in Research - Observational Tools

1.3.2.1 Global Rating Scales

Global rating scales are designed to measure generic skills that can be applied to a wide variety of different settings and procedures (K. Ahmed, et al., 2011). The first well-established global rating scale used in surgical specialties was the Objective Structured Assessment of Technical Skill (OSATS) scale (Martin et al., 1997). The original OSATS scale was devised to assess surgical skill in basic tasks and was piloted in two formats, bench-top models and live animal models. A global rating scale was combined with a task-specific checklist for each skill station (Martin, et al., 1997). Although initially designed to measure basic surgical task performance in six task stations, the global scale component of the OSATS rating scale has found wide adoption across numerous surgical disciplines and assessment formats. For example, the scale has been used in simulated procedures and real operative procedures, either as live ratings or using video review (Aggarwal, Grantcharov, et al., 2007; Dath et al., 2004; Goff, Lentz, Lee, Houmard, & Mandel, 2000; Sarker, Chang, Vincent, & Darzi, 2006). The global rating scale was designed to measure seven relevant skills required during surgeries (Table 1). Three scale items represent technical skill evaluations (*respect for tissue, time and motion, instrument handling*). Four domains reflect competencies such as procedural knowledge, forward planning and use of resources relevant to the task being scored (*knowledge of instruments, knowledge of specific procedure, use of assistance, flow of operation and forward planning*). Each of the seven scale items on the OSATS scale is to be rated on a 5-point Likert scale with anchors described for the lowest, middle and highest score (Martin, et al., 1997).
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Respect for tissue</strong></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 tissues appropriately with minimal damage.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Time and motion</strong></td>
<td>Many unnecessary moves.</td>
<td>Efficient time/motion but some unnecessary moves.</td>
<td>Economy of movement and maximum efficiency.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instrument handling</strong></td>
<td>Repeatedly makes tentative or awkward moves with instruments.</td>
<td>Competent use of instruments although occasionally appeared stiff or awkward.</td>
<td>Fluid moves with instruments and no awkwardness.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Knowledge of Instruments</strong></td>
<td>Frequently asked for 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 required and their names.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Use of assistants</strong></td>
<td>Consistently placed assistants poorly or failed to use assistants.</td>
<td>Good use of assistants most of the time.</td>
<td>Strategically used assistant to the best advantage at all times.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flow of operation and forward planning</strong></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><strong>Knowledge of specific procedure</strong></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>

Overall, on this task, should this candidate: □ Pass □ Fail?

**Table 1: Objective Structured Assessment of Technical Skills (OSATS) global rating scale**

from Martin et al. (1997)

Subsequent scales used to assess performance in numerous disciplines essentially represent modifications of the global component of the original OSATS scale. Modifications include collapsing, augmenting or modifying the scale categories (Beard, Choksy, & Khan, 2007; Birkmeyer et al., 2013; Cremers, Lora, & Ferrufino-Ponce, 2005; Curry et al., 2012; Ezra et al., 2009; Grantcharov et al., 2004; Kumar et al., 2012; Leong et al., 2008; Matsumoto, Hamstra, Radomski, & Cusimano, 2001; J. Park et al., 2007)

A further global rating scale that has been widely used to assess technical skill is the Global Operative Assessment of Laparoscopic Skills (GOALS) scale (Vassiliou et al., 2005). The global rating component was modeled after the global OSATS scale and was designed to assess laparoscopic surgical performance. A 5-point Likert scale with anchors describing the highest, middle and lowest rating was used. The skill categories used for the global rating were “depth perception”, “bimanual dexterity”, “efficiency”, “tissue handling” and “autonomy”. Similar to the OSATS scale, the global scale was also combined with a procedure-specific task list, which was rated on a binary scale (done/ not done) (Vassiliou, et al., 2005). In addition, two visual analogue scales were included to document procedure difficulty and overall competency (Vassiliou, et al., 2005). The GOALS scale has been shown to deliver reliable results for live direct observation as well as for rating recorded procedures (Vassiliou et al., 2007).

Despite these observational global rating scales being termed “objective”, they do require some degree of subjective interpretation. For example, the performance to be rated should only be measured against the criteria or anchors specified in the scale, but raters may be inclined to rate the performance against other criteria such as previously observed behaviors. For example, it has been shown in an experimental setting including internal medicine specialties that borderline performance which was preceded by a poor performance, was assigned higher scores than borderline performance preceded by good performance (Yeates, O’Neill, Mann, & Eva, 2012). This effect, termed contrast bias, needs to be acknowledged when rating performance on global scales. In addition, since global rating scales were designed as generic tools, enabling assessment in multiple different settings, item categories are general and descriptors are mostly vague limiting applications in formative feedback (K. Ahmed, et al., 2011). Checklists, on the other hand, require the observer to specify solely whether a behavior was present or not, making ratings easier for non-expert raters (Regehr, et al., 1998).
1.3.2.2 Task Specific Checklists

Task-specific checklists have frequently been added to global rating scales such as the initial OSATS and GOALS scales (Martin, et al., 1997; Vassiliou, et al., 2005), but they have also been designed to assess procedural performance as stand-alone tools. Checklists, in general, use binary ratings to score performance, where the raters score whether a performance was observed or not. Checklists break down procedures into multiple substeps and tasks that are to be judged individually. Sarker and colleagues (2005) designed a checklist for laparoscopic cholecystectomy that combined a generic skills checklist with an “error checklist” derived from panel, expert discussion and review of the literature. The “Eubanks” checklist represents another well-established task-specific checklist combining general procedure aspects with error metrics designed to score laparoscopic cholecystectomies, with procedure steps derived from expert video analysis (Eubanks et al., 1999). Within the “Eubanks” system, individual aspects of performance are weighted with regards to presumed complexity of steps and severity of errors. Both checklists had been reported to be reliable when expert raters applied the tools in research studies (Eubanks, et al., 1999; Sarker, Chang, Vincent, & Darzi, 2005). In vascular surgery, Wilasrusmee and colleagues (2007) used a procedure-specific checklist for vascular anastomosis creation to identify parameters in a simulation environment that would predict poorer performance in the OR (Wilasrusmee, Lertsithichai, & Kittur, 2007). The checklist used represented a detailed breakdown of steps required to suture an arterio-graft anastomosis.

Most checklists require adherence to a particular order of steps or tasks and do not account for alternate treatment solutions. This step-adherence makes assessments very rigid (Beard, et al., 2007; Zevin, et al., 2013). Furthermore, reports regarding the ability of these tools to discriminate between surgeons of differing training levels have been mixed. Studies have indicated limited discrimination between training levels (Aggarwal, Grantcharov, Moorthy, Milland, & Darzi, 2008; Martin, et al., 1997; Reznick, Regehr, MacRae, Martin, & McCulloch, 1997) and others reported limited discrimination only within an expert group (Beard, et al., 2007). One explanation for this lack of discriminative ability at high skill levels may be that a ceiling effect is reached once a procedure has been learned (Beard, et al., 2007). In view of
lacking assessment methodologies at the higher end of surgical skill, additional metrics such as error counts may prove valuable.

1.3.2.3 Procedure-specific Global Rating Scales

Procedure-specific global rating scales combine the detail of task-specific checklists with the broader scale range of global ratings. The scales are subsequently limited for use in the procedure for which they were designed, but deliver more granular information than global ratings. The design of procedure-specific rating scales is similar to that of task-specific checklists as surgical procedures are broken down into relevant substeps. The rating format though differs as Likert rating scales with descriptive anchors are used to evaluate performance in each substep rather than binary ratings. In the current scales, the process of procedure decomposition has been achieved through by expert video review (Aggarwal, Boza, et al., 2007), expert discussion (Ghaderi et al., 2011; Sarker, Chang, Vincent, et al., 2006), literature review (Sarker, Chang, Vincent, et al., 2006) and expert panel surveys (Palter & Grantcharov, 2012; Zevin et al., 2013). Hierarchical task analysis represents another established method that has been used in the design of rating frameworks. Since the methodology has most commonly been employed in the process of error analysis, hierarchical task analysis will be discussed under the section “error analysis”.

Procedure-specific global rating scales have been designed for numerous procedures including laparoscopic cholecystectomy (Sarker, Chang, Vincent, et al., 2006), colorectal surgery (Dath, et al., 2004; Palter & Grantcharov, 2012), open sigmoid colectomy (Lipman et al., 2010), laparoscopic Nissen fundoplication (Dath, et al., 2004), laparoscopic salpingectomy (Larsen et al., 2008), laparoscopic incisional hernia repair (Ghaderi, et al., 2011), jejunojejunostomy (JJ) (Aggarwal, Boza, et al., 2007), LRYGB (Zevin, et al., 2013) and saphenofemoral junction ligation (Pandey et al., 2006). A variation of procedure-specific rating scales that has been used in vascular surgery is the assessment of an operative outcome as an “end product” scale rather than rating the step execution (Black, Harrison, Horrocks, Pandey, & Wolfe, 2007). Similar to other procedure-specific scales, the scale evaluated surgical skill using a 5-point Likert scale. In contrast to evaluating process the scale evaluated outcomes of the different steps within the task
to be completed such as the quality of the incision, the amount of residual plaque remaining, quality of vessel closure, and quality of suture placement (Black, et al., 2007).

Procedure-specific rating scales in contrast to global rating scales are far more detailed allowing for use in structured formative feedback (Beard, et al., 2007; Larsen, et al., 2008). Subsequently, in a training setting combining global ratings with procedure-specific components has been advocated to enhance the value of formative feedback gained from assessments (K. Ahmed, et al., 2011).

As detailed in the prior section on psychometric properties of tests, when viewing any given assessment designed to measure a particular construct, it is important to explore all validity evidence relevant to that specific construct, as well as the setting in which the tool is to be applied (Downing, 2003). In the context of observational tools used for surgical performance assessment, it is thus necessary to gather evidence from all relevant sources which would include evidence that the content covered by the scale corresponds to the construct of interest, and evidence that the rater’s response process matches what the scale designers intended to be measured. Further, the assessments should be made accurately (reliable) leading to comparable results either between raters (inter-rater reliability) or consistent measurements when one rater uses the assessment over time (intra-rater reliability). Finally, evidence that scores obtained correspond to other measures targeting the same, or similar constructs is also essential. Thus, validity evidence in support of scores obtained through the use of individual assessment tools may not be comprehensive in only one particular study. However, by reviewing score measures obtained in various settings in studies with differing intentions and subject populations a good amount of validity evidence can be summarized in support of measurements made using the most common assessment methodologies described above.

For the example of scores obtained through use of the OSATS scale, several sources of validity evidence can be obtained by examining just a few studies from the surgical education literature reporting validity parameters for OSATS applications. For example, validity evidence gathered in the original OSATS study by Martin et al (1997), established test equivalence for all versions of the scale (global and checklist formats in live and bench-top settings), with moderate to high inter-rater reliability (Cronbach’s alpha 0.61-0.74) for all but the reliability of the scores of the checklist format in bench-top models. Further, strong correlations ($r = 0.81-0.87$) between test
formats (global versus checklist) suggested that similar constructs were measured by the scales (Martin, et al., 1997). Year of training was determined to contribute 29 to 34 percent of score variance, with training year being shown to have a lesser effect on scores measured by the checklist format (Martin, et al., 1997). Pass/fail scores did not show adequate reliability and did not correspond well with training level. The validity evidence for content can be derived from the credentials of the tool designers as experts in the field of surgical education. The response process was reviewed by requesting ten uninvolved surgeons to verify whether the checklist content reflected the practice of the training system in which the assessment was aimed to be made (Martin, et al., 1997). This evidence was subsequently supplemented by a large scale application study that was conducted using only the bench-top format of the OSATS examination (Reznick, et al., 1997). In this study, inter-station reliability was high for both scale formats (Cronbach alpha 0.843 and 0.781 for global and checklist formats respectively) (Reznick, et al., 1997). Furthermore, the level of variability of score results attributed to training level was calculated at 63 percent for the global scale, which was higher than that measured in the original study (Reznick, et al., 1997). Significant differences in scores on the checklist were noted for all training levels, except for differences between postgraduate (PGY) 4 and 5/6 years (Reznick, et al., 1997). The influence of surgical experience on OSATS scores was later confirmed within the context of the OR environment in gynaecology (Hiemstra, Kolkman, Wolterbeek, Trimbos, & Jansen, 2011). Hiemstra et al. (2011) showed that OSATS scores increased by approximately one point per case performed, using a slightly modified version of the global rating scale. Additional validity evidence, for example, was gathered by Aggarwal et al (2007) in the setting of video analysis of laparoscopic cholecystectomies. Significant differences were noted in OSATS scores as measured on the original global rating scale between novice and experienced surgeons (median OSATS score 24 versus 27 respectively, $P = 0.031$). Good inter-rater (Cronbach alpha 0.72) and inter-test reliability (Cronbach alpha 0.72) was determined. Also, significant correlations were found between OSATS scores and scores from motion tracking parameters that were deemed to measure the same construct of surgical skill ($r = 0.4-0.7$) (Aggarwal, Grantcharov, et al., 2007). So to summarize, validity evidence in support of scores obtained through use of the OSATS scales has been gathered in several settings including bench-top, live animals, laparoscopic general and gynaecological surgery, live assessments, and video recordings. The validity of test results obtained by using the OSATS global rating scale has been demonstrated through evidence from content, response process, and internal structure
analysis, as well as by examination of the relationship with other variables. Pass-fail criteria as a form of consequential evidence have not yet been established.

Similarly, as an example of a procedure-specific global rating scale used in this thesis, validity evidence supporting the interpretation of scores obtained through use of the Bariatric Objective Structured Assessment of Technical Skill (BOSATS) can be collated. In the study detailing the design and testing of the BOSATS scale (Zevin, et al., 2013), although reported according to the older framework, several sources of validity supporting score interpretation were explored. A detailed description of how the scale content was devised was included in the report. The item design followed an extensive process. First, a hierarchical task analysis of the procedure of interest (LRYGB) was performed. Hierarchical task analysis was followed by a Delphi expert consensus (Zevin, et al., 2013). The consensus included the opinions of 19 specialists in the field from a wide geographic distribution and aimed to determine which of the items generated in the hierarchical task analysis were deemed relevant for resident training (Zevin, et al., 2013). Test-criterion correlations were established between scores on the scale and level of experience of the surgeon performing the case using volume criteria (Zevin, et al., 2013). Further, correlations with other variables deemed to measure a similar construct (OSATS) were also significant for all subsections of the scale, except for one surgical variation of gastrojejunostomy (GJ) anastomosis creation (Spearman’s Rho ($r_s$) 0.48-0.96, $P = 0.00$). Inter-rater and test-retest reliabilities were excellent for the overall scale (ICC above 0.90) (Zevin, et al., 2013). In a subsequent study, Zevin et al. (2014) showed that by training the skills required for the creation of a JJ, BOSATS scores improved significantly. This finding suggested that the scale measured the same construct as was being targeted with the training intervention (Zevin, 2014). Both studies combined deliver a good amount of validity to support the assumption that BOSATS scores measure surgical skill in LRYGB, with evidence derived from content, internal structure, and relationships to other variables.
1.3.2.4 Error Analysis


1.3.2.4.1 Background

The introduction and wide implementation of laparoscopic surgery have been associated with increased attention to surgical performance and skills assessment. Compared with the open approach, laparoscopic surgery requires the development of a unique set of skills to adapt to the challenges of two- to three-dimensional orientation, the fulcrum effect of the abdominal wall, decreased tactile sensation, and amplification of tremor (Fried & Gill, 2007). At the same time, the ability to document the procedure through the view of the endoscope camera has allowed surgical educators to evaluate surgical performance and technical errors. Learning curves of laparoscopic procedures have been examined in pursuit of designing educational curricula that could lead to proficiency and safe practice (Choi et al., 2009; Dincler et al., 2003; Gould, Garren, & Starling, 2004; I. J. Park, Choi, Lim, Kang, & Jun, 2009).

Most studies to date have focused on the description of technical skill as a measure of surgeon competency for training purposes and as a surrogate measure for safety (Martin, et al., 1997; Regehr, MacRae, Reznick, & Szalay, 1998; Vassiliou, et al., 2005; Vassiliou et al., 2006). However, multiple factors have been identified that may influence patient safety and surgical outcome. These may include the surgical team, social interactions, the technology used, organizational and environmental factors, patient characteristics, and the complexity of the
procedure (Rodrigues, Wever, Dankelman, & Jansen, 2012). Description of errors in the non-technical domain has focused predominantly on the function of the surgical team. Technical and non-technical errors may be associated with intraoperative adverse events and also may influence clinical outcomes (K. Catchpole, 2010; The Joint Commision, 2011). Currently, there is consensus that both technical and non-technical errors should be recognized and prevented with appropriate educational interventions.

Technical error is the most frequently described surgical error type, especially in the evaluation of closed malpractice cases (A. A. Gawande, Thomas, Zinner, & Brennan, 1999; A. A. Gawande, Zinner, Studdert, & Brennan, 2003; S. O. Rogers, Jr. et al., 2006). Technical errors are defined as adverse events directly related to manual errors of the surgeon (e.g., damage to adjacent structures) (Kohn, Corrigan, & Donaldson, 2000). Knowing the nature of technical errors and recognizing potential patterns that lead to adverse outcomes are important steps toward developing rescue mechanisms to prevent adverse outcomes and designing specific training interventions to improve surgical performance. Furthermore, each surgeon should have the opportunity for self-reflection of slips, errors, and adverse events to learn from them (Rebasa et al., 2009).

To date, only a few studies have been designed to describe and evaluate patterns of technical errors. The result is poor understanding and a lack of preventative mechanisms in current clinical practice. This study aimed to identify and review studies describing technical errors in laparoscopic surgery and to analyze the implications of error analysis in this field.

1.3.2.4.2 Methods

Search method

A literature search was performed using the following online databases: Medline (1946-2012 week 14), Cochrane Library, EMBASE (1947-2012, week 14), and OVID (1946-2012, week 14). The search terms “technical Error(s),” “medical error(s),” “technical skill(s),” and “adverse
event(s)” were used in combination with the terms “laparoscopy/laparoscopic surgery” using the Boolean operator AND.

Inclusion criteria

The search was limited to original articles published in the English language in peer-reviewed journals. The search results were screened using article titles and types to exclude articles not relevant to the search. All articles whose titles overtly deviated from the search subject matter as well as duplicates were excluded from the study at this stage. For all remaining articles, the abstracts were screened. Articles providing only an English language abstract but a non-English language full text article and abstracts without any description of technical errors or skills were excluded. Only studies evaluating technical error in the field of laparoscopic general surgery were included in the study.

Procedures in other surgical specialties such as gynaecology were not included in this review because procedure-specific details and intraoperative steps vary greatly between the different surgical specialties. Descriptions of errors associated with open surgical approaches and abstract skill tasks in surgical simulators also were excluded. Opinion papers and reviews were not included in the analysis, but the retrieved reviews were used to identify further articles by cross-reference.

Extracted data and quality of included studies

The articles were reviewed with regard to study aim, operative procedure, study population, method of error description, classification and frequency. Descriptions of technical errors were recorded for comparison. The quality of each study was not analyzed because the review topic focused on error description rather than the design of the underlying study. Nevertheless, strengths and weaknesses of the individual error rating scales such as inter-rater reliability, construct validity, and objectiveness were assessed and discussed.
1.3.2.4.3 Results

The initial search yielded 2,282 articles. Using predefined inclusion and exclusion criteria, the number of articles included in the abstract review was reduced to 137 potentially relevant abstracts. Of these 137 abstracts, 82 were excluded because they had no descriptions of technical skills analysis or errors or because the articles represented reviews, letters, or comments. Of the remaining 55 articles, an additional 34 were excluded because they represented studies describing only technical skills rating scales without an error description, retrospective analyses of adverse outcomes without an error description, or descriptions of open operative skills or abstract skill tasks (errors in virtual reality (VR)). Case reports without error description were also excluded.

A total of 21 articles were subjected for a full-text review. Figure 2 illustrates the search and exclusion algorithm in the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) format (Liberati et al., 2009).
Retrieved studies

The search identified 21 studies that included a specific description of surgical technical error in laparoscopic general surgery (Table 2). Of these 21 studies, 14 (67 percent) were purely observational studies. In 17 studies (81 percent) observations from a real OR environment were presented (Ahlberg et al., 2007; K. Catchpole, Mishra, Handa, & McCulloch, 2008; Eubanks, et al., 1999; Hamad, Brown, & Clavijo-Alvarez, 2007; Hwang et al., 2006; Joice, Hanna, &
Cuschieri, 1998; Lien et al., 2007; McCulloch et al., 2009; Mishra, Catchpole, Dale, & McCulloch, 2008; Sarker, et al., 2005; Sarker, Chang, Vincent, et al., 2006; Seymour et al., 2004; Seymour et al., 2002; Talebpour et al., 2009; Tang, Hanna, Bax, & Cuschieri, 2004; Tang, Hanna, Joice, & Cuschieri, 2004; Way et al., 2003), and four studies (19 percent) consisted of observations obtained from simulation curricula (Adrales et al., 2003; Pugh, Plachta, Auyang, Pryor, & Hungness, 2010; Tang et al., 2006; Tang, Hanna, & Cuschieri, 2005). Two studies (10 percent) were retrospective analyses of documented adverse events (Lien, et al., 2007; Way, et al., 2003), whereas three studies (14 percent) were randomized trials (Ahlberg, et al., 2007; Sarker, Chang, & Vincent, 2006; Seymour, et al., 2002), and two studies (10 percent) were prospective interventional studies (pre/post design) (McCulloch, et al., 2009; Pugh, et al., 2010). A total of 17 studies (81 percent) investigated errors during laparoscopic cholecystectomy. The other procedures used for evaluation in the remaining four studies were laparoscopic pyloromyotomy, JJ LRYGB, gastrojejunostomy (GJ)/ cholecystojjunostomy, and ventral hernia repair (Table 2).
Table 2

<table>
<thead>
<tr>
<th>Reference</th>
<th>Design</th>
<th>Procedure</th>
<th>OR/Box</th>
<th>Evaluation material</th>
<th>Aim of study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Studies investigating error rate and mode in laparoscopic surgery (quality assessment)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joice, et al., 1998</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To identify errors made by surgeons during laparoscopic procedures</td>
</tr>
<tr>
<td>Way, et al., 2003</td>
<td>Retrospective analysis</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To apply human performance concepts to understand the causes of bile duct injury</td>
</tr>
<tr>
<td>Tang, Hanna, Joice, et al., 2004</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To investigate patterns of failure underlying technical errors</td>
</tr>
<tr>
<td>Tang, Hanna, Bax, et al., 2004</td>
<td>Observational study</td>
<td>LPM</td>
<td>OR</td>
<td>Video</td>
<td>To identify technical errors made by surgeons during early experience of LPM</td>
</tr>
<tr>
<td>Seymour, et al., 2004</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To describe a method of surgical error analysis</td>
</tr>
<tr>
<td>Sarker, et al., 2005</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To assess rate of minor and major errors in LC performed by specialists</td>
</tr>
<tr>
<td>Lien, et al., 2007</td>
<td>Retrospective analysis</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To analyze factors leading to CBD Injury</td>
</tr>
<tr>
<td>Mishra, et al., 2008</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Live</td>
<td>To analyze relationship between non-technical skills and technical errors in the OR</td>
</tr>
<tr>
<td><strong>Studies using error descriptions as a surrogate for surgical skill in a training/education setting</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eubanks, et al., 1999</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To develop a method of assessing technical skill during operative procedures</td>
</tr>
<tr>
<td>Seymour, et al., 2002</td>
<td>Randomized, single blinded</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td>To demonstrate that VR training transfers to the OR</td>
</tr>
<tr>
<td>Adrales, et al., 2003</td>
<td>Observational study</td>
<td>Cholangio/ AE/IH</td>
<td>Box</td>
<td>Video</td>
<td>To evaluate construct and face validity of 3 laparoscopic simulation modules</td>
</tr>
<tr>
<td>Tang, et al., 2005</td>
<td>Observational study</td>
<td>LC</td>
<td>Box</td>
<td>Video</td>
<td>To describe technical errors committed by surgical trainees</td>
</tr>
<tr>
<td>Reference</td>
<td>Design</td>
<td>Procedure</td>
<td>OR/Box Evaluation material</td>
<td>Aim of study</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------</td>
<td>-----------</td>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Tang, et al., 2006</td>
<td>Observational study</td>
<td>LC</td>
<td>Box</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To investigate the relationship between OSCE scores and error scores</td>
<td></td>
</tr>
<tr>
<td>Sarker, Chang, &amp; Vincent, 2006</td>
<td>Prospective single blinded</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To develop a structured assessment tool to assess technical and technological</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>skills for laparoscopic procedures</td>
<td></td>
</tr>
<tr>
<td>Hwang, et al., 2006</td>
<td>Observational study</td>
<td>LC</td>
<td>OR</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To evaluate the quantitative measurement of motor performance in LC and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>correlate motor performance with technical errors</td>
<td></td>
</tr>
<tr>
<td>Ahlberg, et al., 2007</td>
<td>Randomized, blinded</td>
<td>LC</td>
<td>Box + OR</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To assess the effect of proficiency-based VR training on outcome of LC in</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>novices</td>
<td></td>
</tr>
<tr>
<td>Hamad, et al., 2007</td>
<td>Observational study</td>
<td>JJ-LRYGB</td>
<td>OR</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To evaluate the effect of videotape debriefing on performance in laparoscopic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>procedures</td>
<td></td>
</tr>
<tr>
<td>Catchpole, et al., 2008</td>
<td>Observational study</td>
<td>CEA/LC</td>
<td>OR</td>
<td>Live</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To analyze the effects of teamwork skills on technical outcome</td>
<td></td>
</tr>
<tr>
<td>Talebpour, et al., 2009</td>
<td>Observational study</td>
<td>Palliative GJ/CJ</td>
<td>OR</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To describe the proficiency-gain curve for palliative laparoscopic GJ and CJ</td>
<td></td>
</tr>
<tr>
<td>McCulloch, et al., 2009</td>
<td>Pre-/post test</td>
<td>CEA/LC</td>
<td>OR</td>
<td>Live</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To assess the effect of teamwork training on error rate and clinical</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>outcome</td>
<td></td>
</tr>
<tr>
<td>Pugh, et al., 2010</td>
<td>Pre-/post test</td>
<td>VHR</td>
<td>Box</td>
<td>Video</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>To investigate changes in operative performance due to feedback focusing on</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>decision making</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2: Overview of selected articles**

Study aims: Assessment of surgical performance by detection and classification of errors in the OR

Eight of the retrieved studies applied error analysis as an approach to investigating surgical performance and determining actual error rates in routine laparoscopic procedures (Joice, et al., 1998; Lien, et al., 2007; Mishra, et al., 2008; Sarker, et al., 2005; Seymour, et al., 2004; Tang, Hanna, Bax, et al., 2004; Tang, Hanna, Joice, et al., 2004; Way, et al., 2003). All of these studies examined errors identified during laparoscopic cholecystectomy. The analyses were conducted not only to quantify the errors, but also to perform root cause analyses of potential causes of error (Lien, et al., 2007; Tang, Hanna, Joice, et al., 2004; Way, et al., 2003)(Table 2 and 3).

Study Aims: Assessment of error as a surrogate for surgical skill in surgical education

The remaining 13 studies assessed technical errors to describe and quantify surgical skill in an educational setting (Adrales, et al., 2003; Ahlberg, et al., 2007; K. Catchpole, et al., 2008; Eubanks, et al., 1999; Hamad, et al., 2007; Hwang, et al., 2006; McCulloch, et al., 2009; Pugh, et al., 2010; Sarker, Chang, Vincent, et al., 2006; Seymour, et al., 2002; Talebpour, et al., 2009; Tang, et al., 2006; Tang, et al., 2005). In these reports, surgical technical error was viewed as a surrogate for skill and was recorded as an independent parameter or in combination with other parameters of surgical skill such as VR metrics, technical skill scales and non-technical skill scores (Table 2 and 4).

Comparison of error scores among different target/ intervention groups showed that error rate was reduced in groups receiving a specific training intervention compared with control subjects (Ahlberg, et al., 2007; Hamad, et al., 2007; McCulloch, et al., 2009; Pugh, et al., 2010; Seymour, et al., 2002). Error analysis also has been used as a method for training with structured feedback and constructive debriefing, leading to an improved surgical performance (Hamad, et al., 2007).
<table>
<thead>
<tr>
<th>Reference</th>
<th>Error categories/ grouping</th>
<th>Additional classification</th>
<th>Error description/ Definition</th>
<th>No. of raters</th>
<th>Results</th>
</tr>
</thead>
</table>
| Joice, et al., 1998        | External error modes      | Corrective action required | Error modes:  
Step not done/ partially completed/ repeated  
Second step done in addition/ instead of first  
Step done out of sequence  
Step done with too much force, speed, depth, distance, time, rotation  
Step done with too little force, speed, depth, distance, time, rotation  
Step done in wrong orientation, direction, point in space  
Step done on/ with wrong object | N/A                        | Identification of performance shaping factors for subtasks  
Identification of 189 errors  
116 intrastep, 73 interstep |
| Lien, et al., 2007         | Patient Environment      | N/A                       | Omission of surgical steps  
Failure to identify anatomy  
Failure to expose Calot's triangle | N/A                        | Incidence of CBD  
Injury reduced to 0 after identification of error modes  
After analysis shorter OR times and significantly less use of drains by attendings |
| Way, et al., 2003          | Misperception Decision/ Knowledge error | N/A                       | Misidentification  
Overshooting incision for cholangiogram  
Poor vision  
Incomplete exposure before clipping  
Incorrect retraction direction | N/A                        | 97% of error were perceptual illusion (misinterpretation)  
Technical faults responsible for 3% of the 252 injuries |
| Tang, Hanna, Bax, et al., 2004 | External error modes | Consequential/ inconsequential | Error modes as described by Joice et al. (1998)  
Consequential:  
Multiple deviated cuts from initial incision  
Splitting in wrong tissue  
Injury to stomach  
Injury to normal pyloric tissue causing bleeding  
Injury to duodenum (no perforation)  
Injury to liver (bleeding) | 1 rater, 2 supervising experts | 310 errors identified  
6 ± 5/ procedure, No correlation between OR time and error frequency |
<table>
<thead>
<tr>
<th>Reference</th>
<th>Error categories/grouping</th>
<th>Additional classification</th>
<th>Error description/Definition</th>
<th>No. of raters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seymour, et al., 2004</td>
<td>Inconsequential: Overshooting instrument movement Nonvisualization instrument tip Instrument movement out of view Applying current without visualizing instrument Avulsion of tissue Cutting without lifting tissues from underlying structures Inappropriate cutting/grasping Insertion of instruments in wrong tissue planes</td>
<td>As described by Seymour et al. (2002)</td>
<td>As described by Seymour et al. (2002)</td>
<td>2</td>
<td>Inter-rater reliability 84-100% Mean rate of error 4.9 ± 1.4</td>
</tr>
<tr>
<td>Sarker, et al., 2005</td>
<td>Minor: Injury to GB (bile spilled) Diathermy liver injury Clip incompletely on cystic artery/duct</td>
<td>Minor/ major Generic/specific Corrected/ uncorrected</td>
<td>2</td>
<td>Inter-rater reliability κ = 0.91 (range 0.80–0.98)</td>
<td></td>
</tr>
<tr>
<td>Reference.</td>
<td>Error categories/grouping</td>
<td>Additional classification</td>
<td>Error description/Definition</td>
<td>No. of raters</td>
<td>Results</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>--------------</td>
<td>---------</td>
</tr>
<tr>
<td>Sarker, et al., 2005 continued</td>
<td>Technical/technological</td>
<td></td>
<td>Minor continued: Misplaced clip fallen into abdomen Cystic artery or branches not identified initially Major: GB Injury (stones spilled) Liver injury with bleeding Unintentional cystic duct division Cystic artery injury Loss of pneumoperitoneum</td>
<td></td>
<td>Significant difference in individual surgical error rate in the minor error group compared to the other groups $P &lt; 0.05.$</td>
</tr>
<tr>
<td>Mishra, et al., 2008</td>
<td>External error modes</td>
<td>N/A</td>
<td>External error modes modified from Joice et al. 1998</td>
<td>1</td>
<td>Mean technical error $2.6 \pm 0.55/\text{procedure}$ Strong negative correlation between technical errors and situational awareness</td>
</tr>
</tbody>
</table>

Table 3: Studies conducted to explore the rate and nature of surgical technical error

[N/A] not applicable; [CBD] common bile duct; [OR] operating room; [GB] gallbladder
### Table 4

<table>
<thead>
<tr>
<th>Reference</th>
<th>Error categories/grouping</th>
<th>Additional classification</th>
<th>Error description/definition</th>
<th>No Raters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>McCulloch, et al., 2009</td>
<td>Technical</td>
<td>Non-Technical Non-Operative Procedural (equipment)</td>
<td>OCHRA approach (Error modes) as described by Joice et al (1998)</td>
<td>2</td>
<td>Effect of training on technical error: 1.73/procedure versus 0.98/ procedure after training ($P= 0.009$)</td>
</tr>
<tr>
<td>Ahlberg, et al., 2007</td>
<td>30-point scoring system</td>
<td>Exposure Clipping Division Dissection errors</td>
<td>Lack of progress, attending takeover Burn/ injury non-target tissue Instrument out of view Inappropriate dissection/ tearing/ division tissue Incorrect angle of GB retraction Dropped retraction Clip overlap/ clip/ spacing error Poor clip application (visualization)/ orientation/ partial closure Nontarget tissue clipped/ clip drop Scissors closure on clip Injury to GB/ cystic duct/ liver Incorrect plane of dissection</td>
<td>2</td>
<td>Dissection errors: 29.5 in control group versus 11.5 VR trained group Clip and tissue division errors: 7.1 in control group, 1.9 in VR trained group Mean errors of exposure: 53.4 in control group versus 15 in VR trained group Frequency of errors in OR: less in VR trained group</td>
</tr>
<tr>
<td>Seymour, et al., 2002</td>
<td>N/A</td>
<td>Attending takeover Instrument out of view Tearing tissue Burn non-target tissue Incorrect dissection plane GB/ Liver injury Lack of progress</td>
<td>GB Injury and burn to non-target tissue 5x more likely in control group (non- VR trained)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eubanks, et al., 1999</td>
<td>Minor/ Major</td>
<td>Point score with weighted errors</td>
<td>Prolonged OR time Injury to other viscus Cystic artery tear Mistaking artery for duct Failure to cannulate cystic duct CBD Injury/ unintentional cystic duct transection Major vascular injury Additional attempt at clip placement/ ductotomy/ misplaced clip Liver injury without bleeding</td>
<td>3</td>
<td>Most common error: liver injury without bleeding (73%) Construct validity good in 2-handed LC method, poor in 1-handed method Frequency of minor errors had greatest discrepancy Inter-observer error-score correlation range 0.55-0.92</td>
</tr>
<tr>
<td>Reference</td>
<td>Error categories/grouping</td>
<td>Additional classification</td>
<td>Error description/definition</td>
<td>No Raters</td>
<td>Results</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Eubanks et al., 1999 continued</td>
<td></td>
<td></td>
<td>Unintentional release of GB</td>
<td>N/A</td>
<td>Novice surgeons could be differentiated from experts based on kinematic measurements Total error scores did not differ between groups: novices average 8 errors/ case, experts average 7 errors/case</td>
</tr>
<tr>
<td>Hwang, et al., 2006</td>
<td>Minor/major</td>
<td>Point score with weighted errors</td>
<td>As described by Eubanks et al. 1999</td>
<td>N/A</td>
<td>Significant higher rate of adverse events in non-debriefed group (26.5% vs 7%) No differences between groups throughout week 4 Minor errors decreased by 29% in debriefed group and 59% in control group (lower overall starting level) from week 1 to week 4</td>
</tr>
<tr>
<td>Hamad, et al., 2007</td>
<td>Adverse events Minor errors</td>
<td></td>
<td>Adverse events: Tears of intestinal serosa or mesentery requiring repair Creation of an ischemic end of the divided jejunum requiring resection Breaking a suture Minor errors: Passing the needle inadvertently through adjacent tissue Twisting suture without completing a knot Inadvertently catching the needle on a loop of suture Dropping tissue or suture</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Catchpole, et al., 2008</td>
<td>External error modes Procedural/ Executional</td>
<td>Errors outside OR field OCHRA approach (error modes) as described by Joice et al. (1998)</td>
<td>2</td>
<td>Mean technical error 1.73/procedure</td>
<td></td>
</tr>
<tr>
<td>Talebpour, et al., 2009</td>
<td>OCHRA classification Others Suturing Electro-surgical Cutting Tissue handling/ manipulation</td>
<td>Bleeding Swording, port dislocation Inadequate depth.spacing of bites Knot tying errors Incorrect needle orientation Needle/suture holding by driver Needle picking by driver Burns/damage with/ without tissue damage</td>
<td>N/A</td>
<td>Errors related to high-frequency electrosurgical hook knife: mean 4.5/case, with tissue damage in 31 instances (34%) Majority of errors occurred during anterior suture line (n=25) with 75 instances of tissue damage, 9 needing repair.</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Error categories/grouping</td>
<td>Additional classification</td>
<td>Error description/definition</td>
<td>No Raters</td>
<td>Results</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>Tablebpour, et al., 2009 continued</td>
<td></td>
<td></td>
<td>Learning curve plateau at 14th anastomosis, reduction in operative time, reduction in technical errors, improved economy of movement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tang, et al., 2006</td>
<td>Procedural/ executional</td>
<td>N/A</td>
<td>Error modes as described by Joice et al. (1998)</td>
<td>1</td>
<td>Correlation between OSCE Scores in skills component and OCHRA, but not in cognitive component Error score as reported in Study from Tang, et al. 2005</td>
</tr>
<tr>
<td>(Tang, et al., 2005)</td>
<td>Procedural/ executional</td>
<td>Consequential/ inconsequential</td>
<td>Error modes as described by Joice et al. (1998)</td>
<td>1 (40 remaining cases)2 (20 first cases)</td>
<td>Highest error probability in clip usage Most frequent: consequential executional errors (62%) 1067 errors identified, 18 ± 10/ procedure</td>
</tr>
<tr>
<td>Adrales, et al., 2003</td>
<td>17-item checklist</td>
<td>N/A</td>
<td>Number of extra tacks placed (IH) Mesh dropped/ position of mesh (IH) Attempts at grasping suture (IH) Attempts at grasping/ clip placement/ inserting catheter (LC) Distal clip attempts (LC) Cystic artery touched (LC) Error in loop ligature (AE) Instrument out of control area (AE/LC/IH)</td>
<td>3</td>
<td>Experience inversely correlated with technical error scores</td>
</tr>
<tr>
<td>Sarker, Chang, &amp; Vincent, 2006</td>
<td>Technical/ technological</td>
<td>Significant major</td>
<td>Significant major: Injury to abdominal viscus Right hepatic artery injury Injury CBD major vascular injury</td>
<td>2</td>
<td>Specific technical skills checklist: no construct validity Generic technical skills checklist: construct validity between novice and expert Good inter-rater reliability (κ = 0.88 - 0.92) for all ratings</td>
</tr>
<tr>
<td>Reference</td>
<td>Error categories/grouping</td>
<td>Additional classification</td>
<td>Error description/definition</td>
<td>No Raters</td>
<td>Results</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Pugh, et al., 2010</td>
<td>Multi-faceted observation from OR steps, errors and decision making</td>
<td>Counter pressure during mesh tacking correct/incorrect Abdominal mesh placement (error &lt; 2cm overlap) Placement of anchoring stitches (error &lt; 4 sutures) Mesh handling correct/incorrect Hernia defect measurement correct/incorrect Tissue handling safe/unsafe Port insertion/visualization correct/incorrect</td>
<td>2</td>
<td>Common errors before intervention included improper visualization of suture passer, improper mesh preparation on back table, omitted anchoring suture. Incomplete hernia repair: 75% before intervention, 0% after intervention</td>
<td></td>
</tr>
</tbody>
</table>

**Table 4: Studies using error descriptions as a surrogate for surgical skill in training/educational setting**

Methods of error analysis: Hierarchical task analysis

Human reliability assessment represents a tool developed to analyze and prevent potential accidents in high-risk industries. A key component in this approach to error analysis is the hierarchical task analysis, a (sub)task analysis of a complex procedure into separate evaluable steps with a focus on the adherence to a predefined order of operative substeps (A. Cuschieri & Tang, 2010).

Hierarchical task analysis has been used to evaluate laparoscopic cholecystectomy (K. Catchpole, et al., 2008; Eubanks, et al., 1999; Hwang, et al., 2006; Joice, et al., 1998; Mishra, et al., 2008; Sarker, Chang, Vincent, et al., 2006; Tang, et al., 2006; Tang, et al., 2005; Tang, Hanna, Joice, et al., 2004); to assess errors performed during palliative gastrojejunostomy (GJ) and cholecystojejunostomy anastomoses (Talebpour, et al., 2009), and to describe errors in paediatric laparoscopic pyloromyotomy (Tang, Hanna, Bax, et al., 2004). This approach represents the most comprehensive error analysis to date. The technical errors observed in these studies could be further classified into interstep (procedural) and intrastep (executional) modes using the external error mode approach (EEM) derived from high risk industries (Embrey, 1986). With this classification, interstep (procedural) errors result from an omission of a specific task, step or from an incorrect step order, whereas intrastep (executional) errors result from a failure to execute a particular individual step correctly. Observed errors also can be classified in terms of their potential consequences as minor or major (Eubanks, et al., 1999; Hwang, et al., 2006; Sarker, et al., 2005; Sarker, Chang, Vincent, et al., 2006), and with regard to the potential requirement for corrective action (consequential/ inconsequential) (K. Catchpole, et al., 2008; Joice, et al., 1998; Mishra, et al., 2008; Tang, Hanna, Bax, et al., 2004; Tang, et al., 2006; Tang, et al., 2005; Tang, Hanna, Joice, et al., 2004).

Methods of error analysis: Human error and accident analyses

Human error and accident analyses use elements of human error and human factors analysis to assess root causes of surgical errors. This approach analyzes not only errors in technical skill, but also errors in knowledge, decision making, perception, as well as external (patient) factors.
Human error and accident analyses also have been used for the retrospective evaluation of known adverse outcomes (Lien, et al., 2007; Way, et al., 2003). In these studies, cases of documented common bile duct (CBD) injuries have been reevaluated with the goal of identifying potentially avoidable root causes of error. Typically, the errors described by this approach consist of situational errors such as poor visualization, incomplete exposure, incorrect retraction, or failure to identify anatomy rather than specific executional errors directly related to dexterity.

**Methods of error analysis: Expert scoring systems**

Several groups have designed error scores for evaluation of technical proficiency in a training environment. These scores, derived from expert consensus based on literature or video assessment, have been used to evaluate study participants in experimental settings, focusing on skill acquisition and assessment of proficiency (Ahlberg, et al., 2007; Hamad, et al., 2007; Pugh, et al., 2010; Seymour, et al., 2004; Seymour, et al., 2002).

In addition, other error tools, such as those described by Sarker et al. (2005, 2006), also have incorporated a grading system to weight observed errors and a measure for error rectification. These scoring systems use a combination of observed events such as injury to gallbladder, liver or cystic duct and error modes such as clip application errors, field of vision errors and dissection errors (Table 3, Table 4).

**Quality of evaluation tools**

The retrieved studies used variable measures to report elements of validity for the respective tools. The data retrieved is summarized in the following section.

**Construct validity**

The ability to discriminate between novice and expert using the Eubanks score was good for two-handed laparoscopic cholecystectomy ($r = 0.50$ and $P = 0.057$) but poor for the one handed
technique described by the authors \((r = 0.02)\) (Eubanks, et al., 1999). Using the same tool, Hwang et al. (2006) showed no differences in scores between novice and expert, but only five individuals were evaluated. Sarker et al. (2006) were able to demonstrate construct validity for a generic checklist used in their study, but not for the specific technical skills checklist.

**Interobserver Reliability**

Using the 100-point weighted score described by Eubanks et al. (1999), reliability was calculated as \(r = 0.55-0.92\), with the greatest discrepancy of recorded frequency observed for minor errors. The 8-point score described by Seymour et al. (2004) had a calculated inter-observer reliability of 84 to 100 percent. The scores suggested by Sarker et al. (2005, 2006) had a high inter-observer reliability for all used checklists, with Kappa coefficients ranging from 0.881 to 0.924 and a \(P\)-value of lower than 0.001.

### 1.3.2.4.4 Discussion

Error training has been suggested as a novel method in surgical education to enhance trainees’ awareness and understanding of errors as well as to devise mechanisms to mitigate the effects of these errors (Darosa & Pugh, 2012). Additionally, it has been hypothesized that ability to identify errors is linked to surgical expertise and therefore should be incorporated in residency training (Bann, Datta, Khan, & Darzi, 2003). Finally, beyond the educational setting, it has been suggested that surgeons should use analysis of errors as a method of self-assessment in routine practice to enhance surgical performance and as a method of quality control (Rebasa, et al., 2009).

This review aimed to identify and analyze all previously published descriptions of technical errors in laparoscopic general surgery and to provide a comprehensive overview of error types, applied rating scales, and potential frameworks for self-assessment to be used by surgical
educators and practicing surgeons. For this review, a heterogeneous group of studies in terms of aims, design, definitions, and outcomes was retrieved.

The reviewed papers used two main approaches to error analysis. The first group aimed to identify, classify, and grade the severity errors committed during operative procedures. Frequently, the objective was to detect and rate errors as a quality control measure or to understand factors leading to adverse outcomes (Joice, et al., 1998; Lien, et al., 2007; Mishra, et al., 2008; Sarker, et al., 2005; Seymour, et al., 2004; Tang, Hanna, Bax, et al., 2004; Tang, Hanna, Joice, et al., 2004; Way, et al., 2003). The remaining studies focused on the use of error analysis for educational purposes (Eubanks, et al., 1999; Sarker, Chang, Vincent, et al., 2006; Talebpour, et al., 2009; Tang, et al., 2005) or as a surrogate measure for skill evaluation (Adrales, et al., 2003; Ahlberg, et al., 2007; K. Catchpole, et al., 2008; Hamad, et al., 2007; Hwang, et al., 2006; McCulloch, et al., 2009; Pugh, et al., 2010; Seymour, et al., 2002). Naturally, in the latter group the actual error description was not part of the main study aim and represented only a component of the methods used to investigate other research questions.

**Approaches to error analysis**

A number of rating scales to assess technical skill for a variety of operative procedures have been developed (Grantcharov, et al., 2004; Martin, et al., 1997; Vassiliou, et al., 2005). Analysis of technical error, however, is distinct from the evaluation of overall technical skill (e.g., dexterity, coordination and tissue respect) as measured in these rating scales. Furthermore, the analysis of errors has been extensively used in the field of VR training for assessment of performance and for constructive feedback. These error metrics, however, are abstract, and although some studies have demonstrated the transfer of training effects from VR to the OR (Kundhal & Grantcharov, 2009), these metrics are not readily transferable into a clinical environment and have limited feedback value. Furthermore, VR metrics also have been questioned regarding their usefulness in training evaluation (Stefanidis, Scott, & Korndorffer, 2009). These different approaches aimed at quantifying skill demonstrate the need for additional objective parameters and tools to rate surgical proficiency.
**Error definitions**

A significant inconsistency in the terminology and definitions of surgical errors was identified in various papers due to the heterogeneity of study goals. To obtain comparability between studies and research groups, the term “error” requires a reproducible definition and an appropriate classification. The Institute of Medicine cites Reason’s definition of an error

> “the failure of a planned action to be completed as intended (i.e. error of execution) or the use of a wrong plan to achieve an aim (i.e. error of planning) (J. T. Reason, 1990)” in the well-known report “To Err is Human” (Kohn, et al., 2000)[pg. 28].

Technical error has been described as an error in which the operative technique has been identified as the underlying cause of an adverse event (Regenbogen et al., 2007). This description includes not only direct manual errors but also cognitive factors such as judgment and knowledge.

Clearly, surgical performance and outcomes are the result of a complex multifactorial process. Success or failure may be determined by a vast number of different elements such as patient-related factors, environmental OR factors, teamwork and team constellation, as well technical equipment performance (K. R. Catchpole et al., 2007; McCulloch, et al., 2009; Mishra, et al., 2008; Rodrigues, et al., 2012). Nevertheless, it seems necessary to attempt to isolate and distinguish elements that reflect surgical technical proficiency by using methods such as skill and error rating. Both of these factors are ultimately related to cognitive and interpersonal skills competencies, but both can be identified in analysis, rated, and used for educational purposes to improve surgeon skill.

**Classification of errors**

Several studies have attempted to classify errors in terms of severity (major versus minor). Their definitions, however, are neither standardized nor objectively quantifiable. One suggested definition describes events with the potential for life-threatening outcomes as major and
individual failures without a serious consequence as minor (de Leval, Carthey, Wright, Farewell, & Reason, 2000).

In the studies reviewed, several groups attempted to grade the observed errors in terms of assumed severity (Eubanks, et al., 1999; Hamad, et al., 2007; Sarker, et al., 2005). Minor errors, for example, included injuries to the gallbladder with bile spilled but no stones spilled (versus major error with stones spilled) or liver injury by diathermy (versus major error with bleeding) (Sarker, et al., 2005; Sarker, Chang, Vincent, et al., 2006). In addition to major errors, “significant major errors” including potentially life-threatening errors such as major vascular injury or CBD injury were defined.

Using these definitions to differentiate between minor and major errors often can be arbitrary and subjective. For example, it is arguable that gallbladder injury with stones spilled will result in a worse outcome than an injury with only bile spilled if treated correctly. On the other hand, the errors labeled as significantly major represent incidents with potential for severe patient injury. Joice et al. (1998) defined errors in terms of whether corrective action is required. Determining an error as consequential or inconsequential may be a better approach to classifying its severity.

Errors, events, and outcomes

In clinical practice, it is important to separate “error” (the undesired action) from “event” (the consequence of this undesired action). A good taxonomy for classification of surgical technical errors is to use the EEM approach (Joice, et al., 1998). Each error described in this classification has the potential to cause inadvertent injury. For example, perforation of the gallbladder during a laparoscopic cholecystectomy could be considered an adverse event with the potential to result in an adverse outcome such as a postoperative abscess. This event could have been caused by a number of different error modes, for example by too much traction with the grasper (EEM “step is done with too much force” (Joice, et al., 1998)) or by poor orientation of the hook diathermy (EEM “step done in wrong orientation” (Joice, et al., 1998)). Equally, factors beyond pure manual execution such as cognitive or interpersonal elements also may lead to the incorrect task execution. However, for analysis of skills and training, objective assessment and detection of observable errors must first be achieved before a more complex root cause analyses is possible.
The previously described generic terminology (Joice, et al., 1998) can be used to assess a vast variety of possible errors. However, the error mode alone may not result in an actual measurable event. Using the laparoscopic cholecystectomy example, the EEM “too little force” applied to the graspers, may only result in accidental slippage from the gallbladder without further consequences. This action fits the previously given description of an error (“the failure of a planned action to be completed as intended” (Kohn, et al., 2000)), but does not result in an actual adverse event for the patient.

Adverse events have been described as unintended incidents in care that may result in adverse outcomes or may require additional care efforts to prevent adverse outcomes (Kohn, et al., 2000). Injuries such as liver or bile duct injuries observed during laparoscopic cholecystectomy (Eubanks, et al., 1999; Sarker, Chang, & Vincent, 2006; Sarker, et al., 2005) should therefore be classified as “events” rather than “errors” because these incidents already represent the outcome of an unintended injury. Applying this terminology, a number of potential error modes can be described, some leading to an adverse event and others not. Observed events, however, must have at least one underlying error mode as mechanism Figure 3.

![Figure 3: Examples of potential error-event outcome relationship](image-url)
Limitations of existing error models

Although the error analysis through hierarchical task analysis using the EEM analysis represents a very comprehensive approach, its use is limited to a research setting. The key component of this evaluation method is the assessment of adherence to a predefined order of operative substeps (A. Cuschieri & Tang, 2010). To use this method, expert consensus regarding acceptable step order is required following procedure deconstruction (A. Cuschieri, 2000; A. Cuschieri & Tang, 2010). Furthermore, specific training in the use of this assessment method is recommended for the evaluation of subsequently obtained data. Therefore, although reliable, hierarchical task analysis applied as a quality assessment method for daily use in general practice may be challenging.

Construct validity and interobserver reliability are important characteristics of a tool. In most of the studies reviewed, more than one observer scored errors, but data for interobserver reliability were not always available. With regards to construct validity, most tools did not offer a good discrimination between expert and novice (Eubanks, et al., 1999; Hwang, et al., 2006; Sarker, Chang, & Vincent, 2006; Sarker, Chang, Vincent, et al., 2006). This may be due to the fact that the majority of studies were performed in an operative environment on actual patients. Consequently, there was no room for major error, and attending surgeons probably would have corrected actions likely to result in major injury.

The “perfect” error tool

To produce a valid, reliable and clinically applicable error assessment tool, the following aspects need to be considered. A broadly applicable definition of what is considered an “error” and what is the resulting “event” is necessary to achieve comparability between groups. The tool should be generic and adaptable to various procedures and should represent a self-explanatory score or checklist to enable wide usage without the need for prior training. It should further have a high interobserver reliability and construct validity, and it should preferably correlate with patient outcomes.
Such a tool could play an important role in surgical education, not only to provide structured feedback and identify areas for improvement, but also to educate residents about technical error and its management. An objective measure of technical error would be a good marker for quality control and credentialing of practicing surgeons, with ultimate potential to have an impact on patient safety.
1.3.3 In-Training Assessment – Workplace-based Assessments

Workplace-based assessments represent the core of competency-based training programs, which have found wide implementation in the United Kingdom (U.K.), Australia and Hong Kong (Singh, Aggarwal, & Darzi, 2014). In contrast to the aforementioned frameworks for assessing surgical skill which have predominately been applied in a research setting, workplace-based assessments aim at evaluating numerous competencies required in the perioperative phase of patient treatment in everyday real life setting. Components of the workplace assessments relevant to the OR are the Direct Observation of Procedural Skills (DOPS) and the Procedure-based Assessments (PBA). DOPS are intended to assess basic technical skill during early training years, aimed to be performed at least on a bimonthly basis (Intercollegiate Surgical Curriculum Programme, 2013). PBAs have been designed for within specialty training and aim to assess technical skills of designated index procedures, as well as certain aspects of non-technical skills. The Intercollegiate Surgical Curriculum Programme (ISCP) in the U.K. advocates that PBA assessments of index procedures should be performed every time the procedure is performed under supervision. Although assessments have been made mandatory for all trainees that entered surgical training since 2007, the time point of when an assessment is to be made is determined by the trainee (Beard, Marriott, Purdie, & Crossley, 2011; Marriott, Purdie, Crossley, & Beard, 2011). The aim of the assessments is to identify learning needs and document progress throughout the course of training (Intercollegiate Surgical Curriculum Programme, 2006-2014). Both assessment scales consist of a checklist format to be rated on the categories “not observed”, “development required” and “satisfactory.” DOPS checklist ratings are intended to be referenced against “satisfactory” performance during the early or core training years (Intercollegiate Surgical Curriculum Programme, 2013). Whereas PBA assessments are to be referenced against the standard expected of a specialist surgeon in practice representing the requirements for the Certification of Completion of Training (Intercollegiate Surgical Curriculum Programme, 2006-2014). In addition, a global assessment completes the evaluation that represents a criterion-referenced rating on a scale from one to four determined by the level of supervision the trainee required to complete the task or procedure. Both assessments also allow for free-text entries and commentary throughout all sections of the form (Intercollegiate
A recent review of assessments completed during a 12-month period between 2008 and 2009 revealed that the mean number of PBAs completed was only 5.4 per trainee and, within a three-month timeframe, a mean of 2.6 DOPS assessments per trainee were completed (Eardley, Bussey, Woodthorpe, Munsch, & Beard, 2013). Besides the scarcity of ratings completed, the results also highlighted that very few low scores had been given suggesting a restricted use of the scale (Eardley, et al., 2013). The authors of the study subsequently pointed out several potential problems that they had identified during the early implementation phase of workplace-based assessments. These problems included the fact that assessments were likely made late in the training phase, deferred by the trainee since the formative intention may have not been fully understood (Eardley, et al., 2013). Also scale use appeared to have been restricted as could be observed through the frequency of central tendency and high score reporting (Eardley, et al., 2013). The average number of PBA forms that had been completed suggested that assessments had not been conducted after every index procedure as intended. A possible explanation for the findings proposed by Eardley and colleagues (2013) was that faculty and trainees may not have been familiar with the concept of formative feedback so that assessments were likely made in a summative rather than formative intent (Eardley, et al., 2013). Although faculty development had been initiated, the staff were not sufficiently trained as evaluators and educators in the early implementation phase. As a consequence of the report, Eardley and colleagues (2013) concluded that the processes around workplace-based assessments had been modified to include better faculty training, promotion of more frequent evaluations, clarification of concepts of formative feedback, and simplification of the rating scales. In contrast, a study conducted in a colorectal department in the U.K. aimed at assessing time requirements for completing PBA forms in the surgical workplace during early implementation of the program revealed, that although PBA form completion required an additional time investment and that restructuring of organizational processes was required to allow for adequate completion of assessments, both trainees and consultants valued the evaluation process which was deemed to generate focused formative feedback enabling the identification of training needs (James, Cross, Lucarotti, Fowler, & Cook, 2009).
Between 2007 and 2009 once the new program had been initiated and implemented a large-scale prospective study was undertaken to evaluate the score validity of the three intraoperative components of PBA assessments (Marriott, et al., 2011). A total of 749 assessments were made during 348 procedures performed by 81 trainees, with 42 percent of assessments made by independent assessors and 57 percent made by clinical supervisors (Marriott, et al., 2011). The results showed that level of specialty training represented the strongest predictor of scores. The amount of variance in scores attributed to the factor "trainee" was determined to range from 33 to 36 percent depending on the scale used (checklist items versus global ratings). Furthermore, the authors calculated generalizability coefficients that were used to model expected reliability measures depending on the number of raters and the number of observations made (Marriott, et al., 2011). Overall the study had gathered strong evidence for the internal structure of the assessment tool as well as evidence through criterion and convergent correlations supporting the validity of measurements made in the context of specialty surgical training in the U.K.

1.3.4 Self-Assessment

Besides external assessments, self-assessments are also an important feature of resident learning. The ability to self-reflect is a core feature of experiential learning as this enables the individual to unmask weaknesses and deficiencies that need to be addressed through more practice or acquisition of further knowledge. In the context of residency training, self-assessment is of high importance especially when external feedback and assessment is lacking. Beyond residency training though, medical practitioners rarely receive peer feedback and must, therefore, commonly rely on self-assessment to identify learning needs. As life-long learning in the form of continuing professional development is a self-regulated process, the practitioner must be able to determine learning needs in order to select appropriate educational activities to fill perceived knowledge and performance gaps (Epstein, Siegel, & Silberman, 2008; Silver, Campbell, Marlow, & Sargeant, 2008). This self-regulation implies that the practitioner will be able to
identify what needs to be learned, but the literature on the topic of self-assessment has shown very inhomogeneous results with regards to the accuracy of physician self-assessments.

Numerous studies in the health sciences have been conducted to investigate the validity of self-assessments by comparing self-rating results to other external measures as reference criteria. A review conducted by Davis et al. (2006) analyzed the quality of these studies to determine how accurately physicians assess themselves. In this review, the majority of studies comparing internal to external assessments failed to show any substantial positive correlation between the two measures (D. A. Davis et al., 2006). The studies reviewed were from a broad medical field including trainees and practitioners, and explored the validity of self-assessments in areas such as clinical skills (Hoppe, Farquhar, Henry, & Stoffelmayr, 1990; Robbins, Kirmayer, Cathebras, Yaffe, & Dworkind, 1994; Tracey, Arroll, Richmond, & Barham, 1997), ability to identify learning needs (Amery & Lapwood, 2004; Tracey, et al., 1997) and technical (procedural) performance (Barnsley et al., 2004; Fox, Ingham Clark, Scotland, & Dacre, 2000; Leopold et al., 2005). Furthermore, Davis et al. (2006) reported, that the inability to adequately self-assess did not appear to be related to the self-assessor's training level or field of expertise (D. A. Davis, et al., 2006). In addition, it has been demonstrated that the individuals that score lowest on external measurements overestimate their abilities and show the least capability to adequately self-reflect (B. Hodges, Regehr, & Martin, 2001; Kruger & Dunning, 1999). A recent review of the capacity to self-assess in the field of technical surgical skills showed slightly more positive findings (Zevin, 2012). In Zevin’s (2012) review, nine of the identified 17 studies demonstrated a positive correlation between internal and external assessments of technical skill (Arora et al., 2011; Brewster et al., 2008; Mandel, Goff, & Lentz, 2005; Moorthy et al., 2006; Nielsen, Foglia, Mandel, & Chow, 2003; Sarker & Delaney, 2011; Sarker, Hutchinson, Chang, Vincent, & Darzi, 2006; Vassiliou et al., 2010; Ward et al., 2003). Despite positive correlations though, self-assessments may still not accurately reflect performance. For example, Mandel et al. (2005) reported high correlations between faculty and trainee ratings, but also showed that the vast majority of trainees had rated themselves lower than the faculty members did. In this setting, a trainee that underestimates performance may lack self-esteem and self-efficacy beliefs which may be counterproductive to learning.

Thus, improving self-assessment accuracy may be a significant step towards ensuring that residents, as self-directed adult learners, can initiate learning in the relevant skill and knowledge
domains. In a review by Michael Gordon (1991), the role of video review as a means to improve self-assessment accuracy and validity was explored in health-care professions. This review showed inconsistent accuracy of self-assessments as measured against external ratings, but did show that video review prompted a more detailed self-analysis although external rating criteria were not adopted as a consequence of the self-review (Gordon, 1991). More recently, a study in medical students also failed to show an improvement in self-assessment accuracy through video review, whilst self-assessments using video review paired with performance feedback metrics led to significantly improved self-assessment accuracy, unguided video review alone did not (Srinivasan, Hauer, Der-Martirosian, Wilkes, & Gesundheit, 2007). As the evidence in support of accurate self-assessment abilities remains mixed, instructional approaches such as facilitated debriefing and feedback will continue to be essential in guiding the learner throughout postgraduate training. Since self-assessment skills do not appear to develop naturally, teaching these skills specifically and implicitly is necessary. Evidence to support the trainability of these skills was also reported in the review by Gordon (1991), with all studies addressing the topic showing improvements in self-assessment abilities as a function of dedicated training. Thus, techniques to improve self-assessment abilities may also need to be included in surgical training.
1.4 Surgical Training in the Operating Room- Instructional Approaches

Surgery is a craft (Traynor, 2011), subsequently in order to become a surgeon, an individual must be skillful in this craft, as is represented by performing successful operations. The skills required for successful operations include technical ability, sound judgment, knowledge and decision-making skills, to name just a few. Although the simulation environment represents an excellent venue to acquire many fundamental aspects of these skills, this environment will never entirely replace the OR. In surgical education, interactions with the senior surgeon as an educator, remains an essential feature (Butvidas, Anderson, Balogh, & Basson, 2011). In addition, teaching within the OR will continue to be a mainstay of training (Canter, 2011; Monkhouse, 2010; Roberts, Brenner, Williams, Kim, & Dunnington, 2012). The OR, however, is an expensive learning environment (Bridges & Diamond, 1999) and subsequently methods used to teach in the OR, need to be scrutinized just as much as those used in simulation to ensure that training is effective (Levinson, Barlin, Altman, & Satin, 2010). Teaching in the OR has largely been described as being unstructured, and the topic remains under-investigated (Iwaszkiewicz, Darosa, & Risucci, 2008). As discussed in the prior section on experiential learning, unguided discovery learning has shown to be inferior to structured or guided approaches. Thus in order to ensure learning in the OR through experience and discovery, instructional strategies must be in place to guide the learner. Therefore, in the following sections, I will detail the main instructional approaches and combination of methods that are commonly used in the OR as well as discuss the relevant evidence base for each of these instructional strategies.

1.4.1 Communication as an educational Strategy

Communication in the OR serves numerous purposes especially during team interactions throughout operative procedures, e.g. to exchange information among team members, and to coordinate team activities. In a training environment though, one of these purposes is to serve as
an instructional approach to facilitate learning interactions between trainer and trainee. Communication interactions will invariably be necessary for all procedures, but not all interactions have an educational intent and only a few can be seen as deliberate teaching behaviours (Roberts, et al., 2012). Even though communication interactions are common and crucial for learning in the OR, only a few studies have focused on identifying communication processes by which the surgeon educates the trainee throughout an operative procedure (Anderson et al., 2013; Blom et al., 2007; Hauge, Wanzek, & Godellas, 2001; Roberts, et al., 2012). Hauge et al. (2001) developed an instrument to classify communication teaching behaviours; Blom et al. (2007) detailed communication types, whilst Roberts et al. (2012) described the purpose of communication processes. Anderson et al. (2013) combined existing knowledge of communication styles with education briefing and debriefing models to enhance intraoperative teaching and sustain the effect over a one-year period.

In order to develop a reliable assessment tool to measure intraoperative teaching behaviours, Hauge et al. (2001), created a framework of four communication behaviours and 26 subcategories of teaching behaviours by conducting observations of general surgical procedures during a two-month timeframe. The four main teaching categories identified were: “instructional”, “questioning”, “responding”, and “setting tone” behaviours (Hauge, et al., 2001). During the observations conducted to determine reliability of the tool, “informing” behaviours were the most common, specifically “directing” communications. “Questioning” behaviours were the least frequently observed, with “asking a closed question” being the most commonly observed behaviour within that category (Hauge, et al., 2001). “Tone setting” behaviours included “commenting”, “joking”, “conversing” as well as negative behaviours such as “chastising”, “insulting” and “making negative jokes”, although negative behaviours remained infrequent (Hauge, et al., 2001). The authors proposed using the tool to determine effective teaching behaviours that could subsequently be used as an evidence base for faculty development initiatives.

Blom et al. (2007) analyzed the content of intraoperative communications to identify teaching processes. This study focused on the extensive analysis of a limited number of observed laparoscopic cholecystectomy procedures to detail communication type and content. Four communication types were described: “explaining”, “commanding”, “questioning”, and “miscellaneous”. The content of the communications was attributed to nine domains including
“operation method”, “anatomy/pathology”, action “location”, action “direction”, “instrument handling”, “visualization” using endoscope, “general”, “private”, and “indefinable” content. Analysis of two full-length procedures revealed an average of 4.3-5.3 communication events per minute. The highest proportion of communication events were defined as “explaining” communications, “commanding” communications were the second most common type. As in the study by Hauge et al. (2001), “questioning” was found to be the least frequent (Blom, et al., 2007; Hauge, et al., 2001). Content analysis revealed that the most explaining communication events involved topics of detailing the operative method, anatomy or pathology, as well as instructions on instrument use (Blom, et al., 2007). Similar to the conclusions drawn from the previously discussed study, Blom and colleagues (2007) proposed that the detailed analysis of communications would enable the creation of targeted training initiatives, albeit the authors suggested using the information to inform curriculum content in contrast to Hauge et al. (2001) who suggested using the information to improve faculty teaching. The rationale given by Blom et al. (2007) was, that since “explaining” communications were frequent, curricula outside the OR likely lacked essential content, leading to a higher need for teaching these contents within the OR (Blom, et al., 2007). This path of reasoning though would infer that the teaching topics chosen by surgeons reflect a lack of trainee knowledge within that content domain, yet surgeons may also elect to discuss topics that they feel are important irrespective of whether the learners already know them or not.

More recently, Roberts et al. (2012) explored intraoperative surgeon-trainee communications to classify the interactions based on educational purpose of the communication. These authors defined four distinct interactions, only two of which had direct educational purpose. “Pure teaching” moments were communications aimed at furthering the trainee's understanding of the procedure or pathology through explanations or demonstrations. The communications did not aim to induce progress within the procedure. Communications designed to move the case forward combined with adequate trainer explanations were equally identified as teaching moments and termed “teaching + instrumental” interactions. On the other hand, the authors also identified communications solely aimed at moving the case forward instrumentalizing the trainee in a directive fashion to complete a task without further explanations. Although these discussions could have served an educational purpose, the trainee was left to deduce the learning content without guidance; as a result, these communications were interpreted as missed learning
opportunities. Lastly, the authors identified communications not directly contributing to the case, termed “banter”, which was thought to contribute to “humanizing” the learning environment (Roberts, et al., 2012). In the cases observed, “pure teaching” and teaching combined with instrumental instructions represented only approximately half of the observed communication interactions (Roberts, et al., 2012). Furthermore, the authors noted that several teaching situations arose from the trainee making a mistake or an error, which led to so-called “teachable moments” (Roberts, et al., 2012). The authors concluded that methods of enhancing teaching and creating an improved structure of interactions in the OR should include goal-setting to focus the communication a priori, intraoperative feedback and post-procedure debriefing (Roberts, et al., 2012).

All three previously discussed studies examined the communication interactions between the surgeon educator and the trainee, each devising slightly differing assessment frameworks, yet commonalities of the findings include the fact that communication, as a powerful educational tool, is not used to its full potential. Although a number of educationally useful communication interactions were described, unproductive interactions were also detailed in all three studies. For example, Blom et al’s (2007) “commanding” appears to correspond with the interactions identified as “instrumental interactions” by Roberts et al. (2012) which direct the trainee without offering further explanation. To improve the educational value of operative experiences, surgeon educators would need to minimize these unproductive interactions, and eliminate negative behaviours such as “chastising” and “insulting” altogether. The knowledge gained through observations should be used as a feedback methodology to improve intraoperative teaching deliberately through modifying communication interactions.

Such a deliberate approach was used effectively by Anderson et al. (2013) in a multistep approach to improving intraoperative teaching. First, teaching practices were assessed through intraoperative observation of communication patterns and teaching behaviours. In this phase, an observational analysis framework differentiating general communication styles from procedural teaching styles was devised. The information gained from the observations was then used as faculty feedback to inform these of their deficiencies. Subsequently, recommendations were made to improve intraoperative teaching. Once the recommended communication and instructional approaches had been implemented, observations were repeated and if behaviours were found to be lacking prompts were given to the faculty members to remind them of the
behaviours that had been agreed upon. These re-observations continued throughout the entire course of the study, although the observation frequencies were reduced (Anderson, et al., 2013). Success of this multi-phase approach was documented through resident surveys conducted at three time-points throughout the study. The targeted intervention improved intraoperative teaching through a reduction of unstructured and ineffective conversation, increasing positive procedural teaching behaviours such as giving feedback, encouraging, and demonstrating (Anderson, et al., 2013). In this study, the authors combined several strategies to improve the trainee’s educational experience. Intraoperative communication interactions were enhanced to focus on those relevant for teaching and combined with preoperative briefing as a means to set individual learning goals for the trainees. In addition, postoperative debriefings were used to consolidate positive behaviours and immediately address deficiencies (Anderson, et al., 2013). This comprehensive approach resulted in improved survey ratings of the index educational behaviours, but whether or not learning was actually improved was not subject of the analysis. Furthermore, in order to induce such a comprehensive modification of educational behaviours, considerable faculty buy-in, as well as significant resources in the form of observers was required. Whether or not the induced change in behaviours could be sustained after termination of the regular observations remains to be determined.

A further very particular form of communication interaction used as an instructional approach in surgery is narrative storytelling to emphasize learning points (Hu, Peyre, Arriaga, Roth, et al., 2012). The learning points addressed by these stories included technical aspects, non-technical aspects, and non-clinical topics (Hu, Peyre, Arriaga, Roth, et al., 2012). Hu et al. (2012) examined the frequency and content of these narrative stories, termed “war stories”. The authors noted that these “war stories” were told to exemplify a teaching point during operative procedures and could be classified into three story types: those detailing “practice changes from lessons learned”, stories describing “personal training experiences”, or stories detailing “adverse events and near misses” (Hu, Peyre, Arriaga, Roth, et al., 2012). The most common teaching points were technical aspects and surgical decision making (Hu, Peyre, Arriaga, Roth, et al., 2012). Error identification and adverse events were also commonly amongst the topics discussed aimed to create awareness and to instruct on future error prevention (Hu, Peyre, Arriaga, Roth, et al., 2012). Although only one in-depth study has been conducted to examine these stories in detail, narrative stories are likely a common phenomenon in the OR environment.
Hu et al. (2012) acknowledge that the narrative stories told may not have been accurate representations of actual events, nevertheless the educational value was deemed of importance despite the lack of an evidence-base underlying the stories. Story-telling is culturally embedded in modern societies, and since the story-telling plays a role in surgical education, more research to explore the educational dynamics around the process is warranted.

1.4.2 Briefing and Debriefing

Briefing in the OR has been identified as a technique to enhance team communication (Lingard et al., 2008), preparing the surgical team for the upcoming case with the aim to improve efficiency, and reduce error (Hicks, Rosen, Hobson, Ko, & Wick, 2014; Neily et al., 2010). In the OR, besides briefing that occurs on a team level prior to the operative procedure, briefing can also occur between the surgeon educator and trainee. In an educational context, briefing prior to field work also aims to promote dialog and orientates the learner, sets goals, and clarifies objectives and expectations (Mackenzie, 2002). This particular form of briefing can aid in establishing clarity on specific learning targets for the upcoming procedure, and thus focus intraoperative learning (Roberts, Williams, Kim, & Dunnington, 2009). Furthermore, a qualitative analysis of residents’ perceptions of instructional interactions showed that residents valued educators that had an instructional plan and that identified specific goals prior to the educational encounter (Vikis, Mihalynuk, Pratt, & Sidhu, 2008). The topic of divergent views on learning goals was explored in a study that used preoperative interviews with trainees and faculty to identify common learning objectives in routine educational procedures (Pernar, Breen, Ashley, & Peyre, 2011). This study revealed a marked divergence with regards to the number of specified learning goals between learners and educators. Trainees defined significantly fewer goals than faculty members did, and several interviewed trainees were not able to specify any preoperative goals at all (Pernar, et al., 2011). Besides the quantity of learning goals, discrepancies were also noted regarding the focus of learning; with faculty identifying goals pertaining to perioperative considerations five times more frequently than trainees (Pernar, et al., 2011). The differences in learning goals and the paucity of goals specified by trainees are of concern as learning opportunities may be missed due to diverging objectives. Briefing, as a
means of promoting dialog, may assist in aligning trainee-trainer expectations. Consciously determining the specific objectives may also aid in resolving disparities in views between trainees and trainers regarding the quality and quantity of intraoperative teaching by ensuring, that both parties are aware of the educational goals and terms of training prior to the case (Butvidas, et al., 2011).

Debriefing has been described as "postexperience analysis" with the purpose of learning from the experience (Lederman, 1984). In experience-based learning, either in a simulation setting or real life in the form of fieldwork, debriefing represents a vital component to enhance the learning experience (Lederman, 1984). Following Kolb’s framework (Kolb, 1984), debriefing can be instrumentalized to strengthen learning in all four experiential learning activities (Walker, 1984). During debriefing, students are required to reflect actively on experiences, try to integrate the experience into pre-existing knowledge frameworks and subsequently apply new concepts to daily practice (Walker, 1984). This reflection should be linked to the primary experience and should be structured around defined learning content (Mackenzie, 2002). In the context of experiential learning, the instructors represent facilitators of learning which can structure the debriefing experience and provide additional input through asking questions that may further stimulate the thought process (Lederman, 1984). Debriefing plays a dominant role in simulation-based training in health-care, and several studies have been conducted to explore the processes to determine effective debriefing techniques and current practices (Arora et al., 2012; Dieckmann, Molin Friis, Lippert, & Ostergaard, 2009; Gururaja, Yang, Paige, & Chauvin, 2008; Raemer et al., 2011). Recently, Arora et al. (2012) conducted an in-depth analysis of debriefing techniques to summarize key components and develop an assessment tool to rate debriefing interactions. Several factors were deemed of relevance to facilitate effective debriefing sessions. These factors included ensuring a safe learning environment, identifying learning objectives, creating an atmosphere of support and avoiding judgment, the use of open-ended questions and techniques of active listening, prompting reflection, guidance in analyzing actions and abilities, and identifying strategies of how to implement the recognized modified behaviours (Arora, et al., 2012).

In simulation, debriefing is frequently facilitated by a trainer and used as a team exercise involving the participants as a group immediately following the practice component (Dieckmann, et al., 2009; Gururaja, et al., 2008). Nonetheless, debriefings can also be conducted effectively as
a pure within-group exercise in team simulations (Boet et al., 2013). Immediate debriefings have the benefit of the experience still being fresh and easily recollected while the ability to structure content or prepare demonstration material may be limited (Gururaja, et al., 2008). Debriefings have also frequently been used after real-life events such as after cardiopulmonary resuscitations (Couper & Perkins, 2013; Edelson et al., 2008). Following these real events, debriefing can also be conducted immediately as “hot” debriefing or delayed as “cold” debriefing (Couper & Perkins, 2013). The benefits of “cold” debriefing following cardiopulmonary resuscitations have been described as the ability to schedule session which may allow for participation of individuals, not primarily involved in the event (Couper & Perkins, 2013). Furthermore, cold debriefing may allow for structuring and preparation of sessions, yet evidence regarding the effectiveness of the technique still remains mixed (Couper & Perkins, 2013). Since substantial variability in reporting debriefing studies has been identified, comparability of results is limited and thus details regarding benefits of particular technical aspects remain unclear (Raemer, et al., 2011).

Debriefing has also been recognized as an important part of teaching surgery in the OR (M. Ahmed, Sevdalis, Vincent, & Arora, 2013; Arora, et al., 2012; Gururaja, et al., 2008; Roberts, et al., 2009). Although debriefing in the OR is common practice as a means of promoting team communication and detecting deficiencies within processes (Hicks, et al., 2014), in an educational context, similar to debriefing in simulations, debriefing represents an instructional strategy that refers to the direct interaction between trainer and trainee following a procedure to reflect on performance and promote learning. Well-structured educational debriefing sessions following an OR encounter remain rare, despite the fact that debriefing has been recognized as a cornerstone of learning from experiences (M. Ahmed, et al., 2013). Although a recent study of debriefing practices revealed that faculty frequently believe they conduct meaningful debriefing sessions, objective review of these practices through direct OR observations could not confirm that effective debriefing routinely occurs (M. Ahmed, et al., 2013). Several factors were proposed as potential barriers to effective debriefing in the OR including time constraints, cultural organizational factors such as the notion of “service” over education, lack of incentive for faculty to invest efforts, barriers in communication, and lack of rapport with the trainees (M. Ahmed, et al., 2013).
Briefing and debriefing activities are closely interlinked as they promote dialog which is essential for learning. Roberts and colleagues (2009) thus proposed a three-phase framework to structure teaching in the OR. This framework included a briefing period to define the learning objectives of the case, followed by episodes of focused intraoperative teaching and concluded by an informal debriefing of the case to promote self-reflection, enhance positive behaviours and correct errors. The proposed framework has yet to be tested with regards to feasibility and efficacy, but nevertheless the model combines the main instructional aspects of experiential learning specifically with regards to feedback and debriefing.

1.4.3 Feedback

Feedback in health-care has been described as:

"an informed, nonevaluative, objective appraisal of performance intended to improve clinical skills" (Ende, 1983)[pg.779]

Feedback is essential in education as it enables the clarification of any divergence between actual and favored performance. Feedback can be divided into the information that is obtained from the individual’s sensory functions, and the information obtained from external sources. Examples of external feedback include performance summaries from a simulator, test scores, or feedback given by an educator. In surgical education in the clinical environment, such as the OR, the primary focus is on feedback derived from external sources, which will be discussed in this chapter. Feedback may have a positive or negative quality, but it should not be judgmental (Ende, 1983). Hence, feedback needs to be differentiated from "unfocused criticism" (K. C. Chung, 2013). The context in which feedback is delivered is of high importance, as to avoid triggering an emotional reaction that may be adverse to the learning process (Ende, 1983). The fear of direct confrontation through feedback has been deemed to lead to grade inflation since faculty have been shown to be reluctant to give poor scores in face-to-face feedback in order to avoid conflict (Colletti, 2000). Beyond the pure avoidance of conflict, the reluctance to give low grades or provide feedback on poor performances may also stem from the fear of subsequent litigation (Kapp, 1981). Therefore, it is essential to understand the terms under which feedback
should be given in order for this powerful educational tool to be effective and in order to avoid negative emotions or defensive behaviour. Jack Ende (1983) summarized the main concepts pertaining to giving feedback through review of feedback practices in health-care and other industries, as well as from the considerations of personal observations. Core concepts included ensuring that the feedback is wanted and expected, based on first hand observations, best referenced against clear goals, given as precise as possible, limited to behaviours that can be modified by the recipient, and delivered in nonevaluative language (Ende, 1983).

Feedback has widely been used as an instructional strategy in health-care. A recent review and meta-analysis of technology-enhanced simulation in health-care education showed that feedback had been used in over a quarter of the studies reviewed (Cook et al., 2013). The meta-analytic review revealed that feedback had a significant positive moderate sized effect on process and product skills (Cook, et al., 2013). The effect was also positive for all other outcome measures although effect sizes were small and remained non-significant (Cook, et al., 2013).

The evidence for feedback in a simulated environment with a focus on procedural skills remains mixed, although the majority of studies explore benefits of feedback in surgical novices targeting only basic skills. Snyder et al. (2009), for example, showed no additional benefit of proctored VR training when compared to independent practice. The target group consisted of medical students performing basic laparoscopic tasks, and VR colonoscopy over a period of eight weeks, with the goal to achieve proficiency as determined by an expert benchmark. The proctored group received feedback from a surgical resident proficient on the simulator (C. W. Snyder, Vandromme, Tyra, & Hawn, 2009). One factor that may have contributed to the lack of benefit of receiving feedback could be the fact that content expertise was deemed equivalent to expertise in teaching and experience in giving feedback. Feedback, as detailed in the prior paragraphs is a skill in itself that does need to be learned. In contrast, Kurglikova et al. (2010) showed that constructive expert feedback with debriefing was highly beneficial to learning endoscopy skills on a colonoscopy simulator. Superior quality of skill acquisition was noted in the group receiving expert feedback compared to the group receiving only feedback generated by the simulator (Kurglikova, Grantcharov, Drewes, & Funch-Jensen). Similarly, a study in gynaecology explored whether medical students would be able to meet expert benchmarks on a VR simulator for the task of a right-sided laparoscopic salpingectomy, and whether standardized feedback would influence the rate at which the predetermined benchmark would be achieved.
(Strandbygaard et al., 2013). The feedback used in this study was scripted and not tailored to the individual’s needs. Nevertheless, participants that received feedback reached the predefined benchmark after fewer repetitions (Strandbygaard, et al., 2013). Standardized feedback, targeting errors made during simulated hand-assisted laparoscopic colectomy was used in a study by Boyle et al. (2011) which combined the feedback with self-assessments and provision of instrument metrics. In this study, the feedback intervention was shown to significantly reduce errors. Thus, the authors emphasized that feedback is a necessary instructional approach to prevent reinforcement of uncorrected errors through practice (Boyle et al., 2011). Standardized, albeit individually tailored, feedback was used in a study by Paschold et al. (2014). These authors examined the effect of targeted individualized feedback on medical students’ performance on basic surgical tasks using a VR simulator (Paschold, Huber, Zeissig, Lang, & Kneist, 2014). The medical students had been categorized as either high performers or low performers based on their baseline (BL) performance profiles. In an attempt to explore whether feedback could be economized, the authors only offered feedback to the low-performing group. The participants in the low-performing group were seen to improve in skill and rapidly match the performance of the high-performing group after the third proctored session (Paschold, et al., 2014). The authors concluded that feedback offered only to underperformers may be useful in an attempt to reduce the amount of faculty input and associated costs (Paschold, et al., 2014).

Similarly, in an effort to reduce costs related to proctoring simulation sessions, Stefanidis et al. (2007) compared the quantity of feedback given to subsequent learning on a basic laparoscopic suturing task. In this study, medical students were either afforded intense feedback, limited feedback or limited feedback combined with video tutorials. The group with intense feedback performed worst, and the group receiving limited feedback and video tutorials performed the best. The authors concluded; that limited feedback was deemed superior to intense feedback (Stefanidis, Korndorffer, Heniford, & Scott, 2007).

It must be noted, that all these studies focus on the role of feedback in a simulated environment, and apart from two studies (Boyle, et al., 2011; Kruglikova, et al., 2010), involved medical students as novice learners. Evidence from the real OR has continued to support the role of feedback as an instructional strategy. For example, operative performance in laparoscopic cholecystectomies was shown to improve significantly after a single session of video-supported feedback was afforded to trainees (Grantcharov, Schulze, & Kristiansen, 2007). Likewise,
feedback combined with video-debriefing performed over a four-week period in complex laparoscopic surgery (LRYGB) was shown to reduce the rate of intraoperative events, although all other metrics (minor error count, and anastomosis duration) were equal between comparison groups (Hamad, et al., 2007). The ACGME general surgery program requirements emphasize the role of feedback in resident education, and state that residents should incorporate formative feedback into daily practice (ACGME, 2012). Furthermore, it has been shown that simply expecting timely feedback may lead to better performance through anticipation (Kettle & Haubl, 2010). Thus, integration of feedback as an instructional approach in the OR seems logical.

Despite recognizing the need for feedback in the OR, routine use of this important instructional strategy is lacking (M. Ahmed, et al., 2013). In neurosurgery, it was noted that although staff and trainees both recognized the need for feedback in clinical education, resident and faculty perceptions of quality and quantity of feedback given in the OR differed considerably (Aoun, El Ahmadieh, Yip, Batjer, & Bendok, 2012). Similar results were reported in obstetrics and gynaecology (Levinson, et al., 2010). These authors showed that residents and faculty agreed that physicians generally give feedback specific to operative steps during cases, but they disagreed on topics of feedback regarding resident surgical skill, instrument handling as well as specificity, amount, and quality of formative feedback (Levinson, et al., 2010).

This high discrepancy between resident and faculty views on how much feedback was given or received was also found in general surgery (Jensen, Wright, Kim, Horvath, & Calhoun, 2012; Sender Liberman, Liberman, Steinert, McLeod, & Meterissian, 2005). In a survey conducted by Sender Liberman et al. (2005) faculty felt they frequently gave feedback, but residents did not agree. Residents and faculty agreed that the content of feedback, regarding aspects that residents wished to receive feedback on was never or only seldom determined before feedback was given. Likewise, in a survey conducted by Jensen et al. (2012) a marked discrepancy between surgical residents’ and faculty’s perceptions of quality and quantity of feedback in the OR was noted. The most obvious and statistically significant difference in perceptions was that while residents did not feel they received immediate feedback after each operation, faculty thought they gave feedback after each case. Similarly, residents rated statements pertaining to amount, specificity, and satisfaction with feedback neutrally on a 7-point Likert scale, while faculty rated these points above the neutral point corresponding with positive agreement, showing significant divergence in perceptions (Jensen, et al., 2012). Possible explanations given for this phenomenon were that
the feedback may have been too informal and took place in passing so that the feedback may not have been recognized as such by the trainee. In addition, lack of opportunities with conflicting obligations in the OR may have also been a cause (Jensen, et al., 2012). Solutions discussed by the authors to help resolve the discrepancies between how faculty and trainees perceive feedback included formalizing the process by using either a written format or dedicating time in a more formal oral session. Another proposed measure was declaring to the trainee before intended feedback that feedback is going to be given (Jensen, et al., 2012).

Besides documenting diverging perceptions of feedback, a lack of regular feedback in the OR was also quantified through direct OR observations (M. Ahmed, et al., 2013). The observations revealed that feedback only occurred in 46 percent of cases, was frequently unspecific, and related primarily to technical skills only. During the observations, trainees rarely initiated feedback through requesting it (M. Ahmed, et al., 2013). Barriers perceived for feedback and debriefing after a case were a lack of debriefing culture and competing clinical duties, whereas complexity of the case and presence of a medical student (competing educational duties) were seen as barriers to intraoperative feedback (M. Ahmed, et al., 2013).

1.4.4 Video-Feedback

Following the theoretical principles of social learning as described by Albert Bandura (1977), people learn through behavioural observations reflecting on outcomes of observed behaviours. While social learning theory mainly draws on principles of learning from observations of others’ behaviours, self-observation has also been shown to be an effective stimulus in problem solving and in the acquisition of motor skills (Ferrari, 1996; N. J. Hodges, Chua, & Franks, 2003). Beyond the traditional behaviour modeling theories it has been advocated that behaviour modeling of motor performance entails two components: observation of the model which is to be imitated, and self-observation in order facilitate regulation of performance patterns to mimic the observed model (Ferrari, 1996). Video-feedback, as a technology-enhanced form of self-observation, has been shown to improve learning of complex motor skills (N. J. Hodges, et al., 2003). In sports training, video feedback represents an established instructional strategy
that has successfully been used in various sports including gymnastics (Boyer, Miltenberger, Batsche, & Fogel, 2009), swimming (Hazan, Johnstone, Martin, & Srikanthswaran, 1990), golfing (Guadagnoli, Holcomb, & Davis, 2002), and tennis (D. Scott, Scott, & Howe, 1998), to name just a few.

In surgery, video playback of trauma bay resuscitations has shown to improve efficiency (Hoyt et al., 1988), and has been shown to lead to superior improvements in team behaviours when compared to feedback alone (Scherer, Chang, Meredith, & Battistella, 2003). The routine use of video-recording in the OR to enable performance analysis has been recommended as a powerful adjunct to routine workplace-based assessments (Beard, 2008). Still, video playback has not found wide adoption outside the context of trauma bay resuscitations. This lack of adoption, may be due to inconclusive evidence regarding the effectiveness of video-feedback in surgical technical skills development. An explanation for the mixed results obtained in studies exploring the effectiveness of video playback may be the variations of how video-feedback was used to supplement training, and variability in learners or skills targeted. For example, video playback can be used in a self-directed fashion without mentoring or guided feedback (Farquharson, Cresswell, Beard, & Chan, 2013; Jamshidi, LaMasters, Eisenberg, Duh, & Curet, 2009), or can be employed in conjunction with expert feedback (Backstein, Agnidis, Sadhu, & MacRae, 2005; Hamad, et al., 2007; Nakada, Hedican, Bishoff, Shichman, & Wolf, 2004). Furthermore, video-feedback can include only playback of the individual’s performance (Backstein, Agnidis, Regeh, & Reznick, 2004; Farquharson, et al., 2013; Hamad, et al., 2007; Nakada, et al., 2004), or can be combined with examples of good and poor performance or expert comparisons (Lin et al., 2003), as is common practice in sports training (Hazan, et al., 1990; Liebermann, et al., 2002). In the sports context, Liebermann et al. (2002) emphasized that the amount of information available to the learner through the video review may be overwhelming and suggested that the technique should be used in combination with mentored guidance especially in novice learners (Liebermann, et al., 2002).

In surgery, two studies exploring the effectiveness of video playback as an instructional method, either as a self-debriefing tool (without external feedback) or in conjunction with external expert feedback (as augmented feedback) failed to show significant effects as measured on performance metrics of global rating scales (Backstein, et al., 2004; Backstein, et al., 2005). Both studies explored skill acquisition in bench-top models. The first study explored skill acquisition in
orthopedic skills across resident levels with only one session of feedback per task alternating instructional approaches at each bench-top station from no feedback, self-debriefing using video playback, and expert facilitated video feedback (Backstein, et al., 2004). The second study, by the same research group, aimed at instructing first-year junior residents the complex task of vascular anastomosis creation using repeated video feedback facilitated by an expert over three weekly practice sessions (Backstein, et al., 2005). Both studies were not powered to detect changes on the scales selected for evaluation. The authors noted that the improvements measured were small and that the global rating tools may have lacked the sensitivity required to adequately reflect improvements. Although no significant results were observed, qualitative analysis of participants’ views on video feedback showed that the participating residents did see a benefit of the instructional approach. Nonetheless, residents also criticized an inconsistency of expert feedback as given between weeks suggesting that the experts were not sufficiently trained in the methodology of video facilitated feedback, and may have been progressing through their own learning curve during the study (Backstein, et al., 2005). In contrast, a study focusing on teaching a basic laparoscopic suturing task showed that video review of own performance significantly increased the individual's ability to perform the task as compared to performance of individuals that did not have an opportunity to review their own performance (Jamshidi, et al., 2009). In this study, the video review was conducted twice at home by the participants without further feedback or guidance (Jamshidi, et al., 2009).

In an OR environment, Hamad et al. (2007) used video debriefing as an instructional approach to improving technical skills in complex laparoscopic surgery (Hamad, et al., 2007). In this non-randomized study, residents performed LRYGB procedures under staff supervision during their minimally invasive surgery rotation. The first cohort of six residents did not undergo debriefing whilst the last cohort of six residents underwent video review together with the attending surgeon. The authors showed that although operative time decreased in both groups throughout the rotation, no differences in minor error rate between groups was noted, while the rate of adverse events was observed to be higher in the non-debriefed group (Hamad, et al., 2007).

A further training model targeting residents that also used video feedback and video modeling in the OR was reported by Lin et al. (2003) in laparoscopic colorectal surgery. The training model consisted of simulation-based training in animal models followed by structured training in the OR. The training in the OR included preoperative tutorials and postoperative debriefings using
video review and feedback, with a focus on error analysis to correct observed mistakes. The residents also had an opportunity to engage in video modeling by reviewing archived videos of cases performed by their mentors (Lin, et al., 2003). The reported data included a three-year period during which the new structured curriculum had been implemented. No significant differences in procedural durations or patient outcome data were observed when compared to an historic BL during which only faculty members performed the cases. In addition the residents’ reactions to the curriculum were positive (Lin, et al., 2003).

In contrast to using video playback to improve specific surgical tasks directly, video review has also been utilized in a more holistic fashion as a means to facilitate surgical coaching in practicing surgeons (Hu, Peyre, Arriaga, Osteen, et al., 2012). In a qualitative study using a grounded theory approach four complex operations were reviewed using video playback (Hu, Peyre, Arriaga, Osteen, et al., 2012). The videos were reviewed by the surgeon performing the case together with a “coach” that facilitated the video-review and self-reflection process. The analysis prompted discussion on technical aspects of the procedure but also included non-technical aspects such as decision-making. The method was deemed to be a valuable source of continuing professional development and was well received by the participants (Hu, Peyre, Arriaga, Osteen, et al., 2012). Similarly, positive reactions to video-debriefing were reported by Nakada et al. (2004). The authors noted that the majority of practicing urologists that had undergone expert facilitated video-feedback as part of a two-day laparoscopic skills course had rated the videotape critique session as “excellent” (Nakada, et al., 2004). The video-debriefing intervention, though, was part of a skills course which included bench-top tasks and a porcine lab. Therefore, inferences from performance improvements cannot be made.

1.4.5 Coaching and Mentoring

Coaching is a process aimed at inducing change on a behavioural, cognitive or emotional level (Ives, 2008). Concepts of coaching originate from performance enhancing models used in sports (O’Connell, Palmer, & Williams, 2012) and have been adapted to be used in numerous contexts including career, relationship, and general life coaching (Ives, 2008). Depending on the
underlying psychological constructs or educational principles, a wide variety of different coaching approaches have been described and used. Although many details differ according to the approach, common principles include the individualized nature of the approach, the aim to induce change, the central role of listening and questioning techniques, the core belief of empowering the coachee to induce the desired change within him/herself (Ives, 2008).

Since coaching techniques are applied to induce change, enhance performance or modify beliefs, coaching needs to be differentiated from other techniques such as counseling, mentoring and training (O'Connell, et al., 2012). The distinction between coaching and counseling is that coaching is not a therapeutic approach, and it is to be used without clinical goals (O'Connell, et al., 2012). Training differs from coaching as the trainer aims to transfer knowledge in a directive fashion adhering to predetermined content (syllabus), whereas coaching is viewed largely as a non-directive process intended to facilitate change rather than teaching (O'Connell, et al., 2012). Lastly, mentoring and coaching need to be seen as distinct techniques, although in health-care the terms have been used interchangeably, and approaches have been combined (Briet et al., 2010).

Mentoring generally refers to a longer process (without a predetermined time-frame), usually as a relationship between a more experienced mentor, guiding the mentee with advice and knowledge to ensure that long-term, mostly professional goals are met (Buddeberg-Fischer & Herta, 2006; O'Connell, et al., 2012; Sambunjak, Straus, & Marusic, 2006). Mentoring has been used in surgery as a means to transmit new knowledge, or support career development and advancement (Briet, et al., 2010; Buddeberg-Fischer & Herta, 2006; Gagliardi & Wright, 2010; Sambunjak, et al., 2006). For example, Briet et al. (2010) described the process surrounding the implementation phase of a novel complex laparoscopic procedure that involved a mentor-mentee relationship in the initial phase of procedure implementation. The authors described that visiting expert gynaecologic surgeons assumed the role of mentors and were available to guide the local practicing surgeons (mentees) through the cases during the early learning stages (Briet, et al., 2010). With the progression of skill and knowledge, the initially guided relationship became less directive until the mentor determined that the mentee was competent to proceed without further guidance. The aim of the process was to make the implementation phase of a novel procedure safer and to ensure that surgeons reached a specified predefined “competent” level prior to independent practice (Briet, et al., 2010). Similar programs have also been described in general
surgery, as a means to ensure that specific (novel) skills are acquired by trained practicing surgeons (Gagliardi & Wright, 2010).

Coaching in contrast to most mentoring relationships represents an agreement between the coach and coachee over a defined time-frame with clearly set goals that are to be achieved within this relationship. In contrast to mentoring, the coach does not necessarily need to be a content expert, as most coaching strategies rely on a rather non-directive approach which aims to guide the coachee towards a solution through questioning and inducing reflection, rather than giving direct advice (Ives, 2008). Since coaching may serve multiple purposes, a number of diverging concepts have been used including humanist, behaviourist, cognitive, adult development, adult learning, systemic, positive psychology and goal-focused approaches (Ives, 2008). Whereas the most approaches target changing personal beliefs and thought processes, goal-oriented approaches focus on specific aims and goals that are to be addressed in a relatively short time-span (Ives, 2008). This focused, and goal driven, approach seeks to use the individual’s own resources to seek out solutions to the problems that have been defined (O’Connell, et al., 2012). These goal-oriented, solution-focused approaches have recently been applied to a surgical context with the aim of improving technical performance. Within solution-focused coaching, a number of variations of session structures have been described, while the fundamental principles of the approaches remain constant (Grant, 2011). Although evidence regarding the effectiveness of any specific coaching structure is still lacking, evidence from therapy applications in psychology suggests that session structure may be useful in slow-moving sessions or in aiding novice coaches to manage the fluid coaching process (Grant, 2011). Since coaching is a tailored approach, the type of structure chosen depends not only on the coach’s preferences, but also on the coachee’s personal needs, as well as context and stage of the coaching process. Thus, structure adherence represents a continuum from minimal structure to well-defined mapped out sessions, with the level of structure shifting throughout the coaching process (Grant, 2011). In numerous solution-focused coaching approaches, the structures follow steps and have been built around acronyms to facilitate use. For example, a seven-step solution-focused framework termed the “PRACTICE” model defines the coaching structure around defining the Purpose of coaching program (or Problem identification/Preferred Outcome/Preferred Options); setting Realistic relevant goals, deriving Alternative solutions, subsequently Considering consequences of the solutions, Targeting most feasible solutions, Implementing Chosen solutions, and Evaluating
implemented solutions (O’Connell, et al., 2012). The PRACTICE model has been modified to accommodate variations in coaching approaches and contexts; thus, several starting points have been included in the first step while still maintaining the original acronym as the “P” can stand for several processes (Palmer, 2011). A further acronym that has been widely incorporated into a number of solution-focused coaching approaches is the development of “SMART” goals which stands for **Specific, Measurable, Achievable, Relevant and Time-bound** (O’Connell, et al., 2012). The identification of SMART goals represents an essential step in solution-focused goal oriented approaches that drive the subsequent coaching process (O’Connell, et al., 2012). In surgery, a recent study exploring improvement of performance in laparoscopic cholecystectomy through coaching used the “GROW” model (Singh, Aggarwal, Tahir, Pucher, & Darzi, 2014). Within the GROW structure, coaching follows four separate phases: **Goal** setting, assess **Reality**, identify **Options**, **Wrap-up** the session identifying next steps (Grant, 2011). The authors adhered closely to the framework, using a set of predetermined key questions to guide the coaching process and augmented the sessions through video-feedback (Singh, Aggarwal, Tahir, et al., 2014). This structured coaching approach significantly enhanced task performance in a simulated environment in comparison to training without additional coaching (Singh, Aggarwal, Tahir, et al., 2014). A further study in a simulated environment that equally showed significant skill enhancement used a three-step coaching model to enhance surgical performance in simulated cholecystectomy procedures (Cole, Mackenzie, Ha, Hanna, & Miskovic, 2014). Cole et al. (2014) used the steps: **set** learning aims before the simulated procedure, structure **dialogue** as coaching throughout the procedure and summarized the session in a feedback **closure** step (Cole, et al., 2014). This study thus used concurrent coaching during the actual task execution in contrast to the aforementioned study that used dedicated coaching sessions after each procedure (Cole, et al., 2014; Singh, et al., 2014). The process detailed by Cole et al. (2014), used a more directive approach, transferring knowledge through expert guidance as a result, the approach resembles training more closely than coaching, as coaching aims to stimulate change from within, facilitated by the coach in less directive fashion. Both studies focused on teaching a particular skill to novice learners, while Greenberg et al. (2014) in contrast proposed a coaching framework to promote continuing professional development, and performance enhancement in practicing surgeons (Greenberg, Ghousseini, Pavuluri Quamme, Beasley, & Wiegmann, 2014). The framework the authors suggested was based on three separate coaching activities: **setting goals, encouraging and motivating, and developing and guiding** (Greenberg, et al., 2014). The
proposed framework remained flexible, in order to accommodate a wide variety of coaching contexts, coachee profiles, and coaches abilities and preferences (Greenberg, et al., 2014).

1.4.6 Task Deconstruction - Graded Responsibility

In medical education, trainees are expected to learn their chosen discipline over the course of dedicated training years. As a trainee passes through the training program, responsibilities shift as abilities increase. Ideally, the learner will come out of the educative process as an autonomous, well-trained independent practitioner. This gradual transition from observer to active participant and independent surgeon relies on principles of graded responsibility, a characteristic feature of apprenticeship style education. Graded responsibility, refers to the concept that a trainer transfers responsibility to the learner as suitable for the level of training and as deemed suitable by the educator as the expert. In the U.S., graded responsibility has been recognized as a core paradigm and has been explicitly detailed by the regulating body in the program requirements for general surgery (ACGME, 2012).

Since, completing operative procedures forms the heart of the surgical disciplines it is the ability to perform surgical procedures safely that is the major focus of training. Clearly, novice surgeons will not be able to perform operative procedures from the first day they enter the program; rather, these skills will progressively be learned. As to allow trainees to participate actively, operative procedures need to be deconstructed into smaller steps or tasks. This approach to teaching surgery, based on graded responsibility, is referred to as task decomposition. In many instances, task decomposition occurs unwittingly, without deliberate structure, and is based on the experience of the educator. Recent publications though have reviewed the method in a more formal fashion. The authors of these publications have drafted systematic approaches to teaching complex procedures (Ali, Rasmussen, & BhaskerRao, 2007; Hashimoto et al., 2012; Marangoni, Morris-Stiff, Deshmukh, Hakeem, & Smith, 2012; Rashid et al., 2006). Task deconstruction allows trainees to participate even in advanced cases, gathering experience in a guided supervised fashion, while ensuring patient safety and OR economy by limiting procedure duration (Ali, et al., 2007; Marangoni, et al., 2012).
Another educational model based on principles of graded responsibility that has been developed to teach in the OR has been termed the “Zwisch Model” (DaRosa et al., 2013). This training model, named after the originator Joseph Zwischenberger, defines four distinct levels of active participation during a case that a trainee must pass through prior to being deemed competent to perform the case independently. In contrast to task decomposition, the model intends the trainee to participate with an increasing level of autonomy throughout an entire case, rather than just completing steps. The four stages described are “show and tell”, where the trainee plays the role of an observer in the first assist role, “smart help”, which entails the trainee assuming the first surgeon role but with active supportive assistance by the faculty, “dumb help”, where faculty involvement is reduced to passive support, and finally the “no help” phase, where the faculty may opt to observe unscrubbed and to guide only when needed (DaRosa, et al., 2013). The model was also proposed as a means to document trainee performance and progress throughout rotations by examining how long individuals remain in any given phase before progressing. The model lends structure to the otherwise unstructured teaching in the OR and could naturally also be combined with other instructional methodologies such as feedback and debriefing, but implementation and feasibility still need to be evaluated. Furthermore, the structure of this method implies that a particular case would need to be common enough to allow repetitive exposure in each of the four stages. Subsequently, this method may be limited to only routine learning cases. Another important factor to consider in the context of graded responsibility is that it requires expert judgment of when to transfer responsibility to the trainee. This decision–making process may remain hard to define and remains influenced by subjectivity (ten Cate, 2006).
1.4.7 Error Focused Training Curricula

Human error is unavoidable, and surgeons of all training levels are prone to committing errors. Furthermore, numerous studies have highlighted that technical errors during surgery are frequent (Ahlberg, et al., 2007; Sarker, et al., 2005; Talebpour, et al., 2009; Tang, et al., 2005; Tang, Hanna, Joice, et al., 2004). In the context of experiential learning, errors may represent a good learning opportunity, provided there is a system in place that enables the identification and interpretation of these errors to promote learning (Dror, 2011; Roberts, et al., 2012; Wu, Folkman, McPhee, & Lo, 1991). In the context of postgraduate surgical education, trainees receive very little instruction about what constitutes an error, which may be because systematic approaches to identifying and reducing errors are lacking (Satava, 2005). Richard Satava (2005) noted that, in the surgical community, the focus has been on complications, which may be because systematic approaches to identifying and reducing errors are lacking (Satava, 2005). In addition, faculty supervision and prompt faculty error mitigation may hinder trainees in gaining adequate experience in recognizing and managing errors during their training (Darosa & Pugh, 2012). Therefore, a number of educators have advocated the introduction of “error training” with a focus on identification, mitigation, and development of error recovery strategies (Brannick, Fabri, Zayas-Castro, & Bryant, 2009; Darosa & Pugh, 2012; Dror, 2011; Kohls-Gatzoulis, Regehr, & Hutchison, 2004; Meyerson et al., 2012; D. A. Rogers, Regehr, & MacDonald, 2002; Satava, 2005).

Within an error-focused training, the first step is to define errors so that a learner will recognize them (Dror, 2011; Meyerson, et al., 2012; Satava, 2005). The ability to detect errors has been shown to be linked to surgical experience and technical skill (Bann, et al., 2003; Bann, Khan, Datta, & Darzi, 2005), as well as to surgical cognitive skill (Kohls-Gatzoulis, et al., 2004). Bann et al. (2003) developed synthetic models of surgical end-products such as skin closure, bowel enterotomy closure, small bowel anastomoses, and arteriotomy closures which contained errors that were to be detected (Bann, et al., 2003). The authors showed that with increasing surgical seniority trainees were more likely to detect errors demonstrating a more profound understanding of correct task execution (Bann, et al., 2003). In a second study, the same authors repeated the error-recognition test but combined the results with global skills ratings on a modified OSATS.
rating scale (Bann, et al., 2005). The authors showed that higher skill scores correlated with higher scores on the error-detection test (Bann, et al., 2005). Thus, with growing surgical experience, trainees not only become more skilled but also the ability to identify incorrect task execution increases. This ability to critically review performance is an essential skill required to determine learning needs in training and continuing professional development.

A study in orthopaedics also used an error detection test, although the authors used the test as a means to assess learning in a course teaching total knee arthroplasty (Kohls-Gatzoulis, et al., 2004). During the arthroplasty course, the authors showed that trainees that had received designated cognitive skills training despite having completed fewer practical repetitions performed comparably, with regards to technical skills, to trainees that had received only technical skills training (Kohls-Gatzoulis, et al., 2004). In addition, the cognitive training group outperformed the skills training group in the ability to detect errors on end-product analysis (Kohls-Gatzoulis, et al., 2004). This study thus demonstrated that investing curriculum time to increase awareness about correct task execution at the cost of technical skills training time did not have an adverse effect on skill acquisition. Also, this cognitive training increased the learners’ ability to differentiate between good and poor performance, increasing the trainees’ ability to critically reflect on performance. This ability to reflect critically and review performance is an important aspect of surgical skills training as repeating tasks without reflection may lead to reinforcement of erroneous task execution which may be hard to remediate.

Although the notion of an error-focused curriculum infers a holistic view of surgical performance and error awareness, studies have also been conducted solely focusing on using error instruction to improve technical skill (Bingener et al., 2008; D. A. Rogers, et al., 2002). These studies applied behaviour modeling techniques using examples of erroneous task execution as a means to improve surgical skill with varying effectiveness. In one study, trainees were instructed on how to perform the basic task of surgical knot tying using either no video, or example videos containing no errors, errors, or a mix of both. The best skill acquisition in this study was achieved by demonstrating the videos containing correct and erroneous examples (D. A. Rogers, et al., 2002). A similar study explored the effect of a teaching video, explaining common errors in laparoscopic suturing, on skill acquisition of the basic surgical task of laparoscopic suturing (Bingener, et al., 2008). The authors showed that the group of medical students that had received the error video performed subsequent suturing tasks slower than the
group that had not seen the video, although both groups improved significantly from pre- to post-test. The quality of task execution was assessed by OSATS global rating scale and showed no differences between the groups. On the error-detection test that was also conducted, the error-trained group outperformed the control group (Bingener, et al., 2008). The authors concluded that the error training had a negative impact on skill acquisition. It must be acknowledged though, that the exercise chosen for analysis is deemed one of the hardest basic tasks to learn. Furthermore, the study design only allowed for 30 minutes of practice. For this particular task though, it has been shown that medical students required a mean of 28 practice repetitions to reach proficiency benchmarks (D. J. Scott et al., 2008). The students participating in the study by Bingener et al. (2008) though, would not have been able to perform more than five repetitions in the time period of 30 minutes based on the best average post-test time score. It is therefore likely that error training, especially in novice learners, may initially result in more mindful practice slowing down individuals, but this could well lead to improved skills at later stages of the learning curve which were not investigated in the aforementioned study.

Overall, in the current light of increasing patient safety, including topics pertaining to surgical error in surgical residency curricula appears necessary to ensure that trainees are aware of potential risks. Although story telling does address some issues relating to error education during intraoperative teaching, this is done in a less deliberate fashion. Cautionary “war stories” may aid the learner in understanding surgical risks; but, they cannot replace a deliberate explicit error education. Deliberate error education is needed to ensure trainees understand surgical errors, recognize these, and learn how to deal with them prior to transitioning into independent practice.
1.5 Limitations of Learning in the Operating Room

The OR has been acknowledged as a central venue for resident learning. Numerous instructional strategies and educational approaches have been shown to be effective through structuring the discovery nature of learning in the OR. Still, a number of barriers can be identified limiting how the OR is used in postgraduate education. In the next section, I will address the most relevant obstacles and limitations to using the OR as a classroom. Specifically, barriers due to organizational or social factors as well as barriers due to the influence of the surgeon as an educator and the residents themselves need to be discussed.

1.5.1 Organizational Factors - Impact of Work-Hour Restrictions

The time available for surgical training has been reduced over the last 20 years, especially as a result of implemented work-hour directives. Cohorts of residents trained 20 years ago completed twice as many training hours than today’s residents (Traynor, 2011). Thus, the impact of work-hour restrictions has been assumed to effect surgical training adversely since surgical technical, as well as decision making skills, are predominantly acquired in the OR, through direct observation and participation in procedures. The time that a trainee spends within the OR is limited. Raphael Chung (2005) quantified the estimated time that a trainee in the U.S. would have spent in the OR throughout their entire five-year training period, with percentage calculations based on an 80-hour work week. The results of this calculation were alarming, with just over 15 percent of training time spent in the OR assisting or operating (R. S. Chung, 2005). The numbers used to calculate the time spent operating were derived from surgical operative logs and included a timeframe between 1998 and 2003 (R. S. Chung, 2005). These statistics thus, reflected caseloads for residents completing training before implementation of the reduced work-hours suggesting that the actual percentage of time spent in the OR may be lower. Although the introduction of reduced work-hours was assumed to lead to reduced caseloads, the evidence gathered so far is inconsistent. A recent review on the topic highlighted that the impact of duty
hour restrictions on operative exposure varies depending on the operative specialty, the case type as well as the level of active surgical participation (first surgeon, teaching surgeon or first assistant) (Jamal, Wong, & Whalen, 2014). A study from the Michigan State University General Surgery training program revealed that total caseload and “major” cases performed by graduating level residents (chief residents) declined after implementing the work-hour restrictions, although all case numbers remained above the minimum requirements as established by the local resident review committee (Damadi, Davis, Saxe, & Apelgren, 2007). Similarly, a further study in the U.S. at Carolinas Medical Centre revealed a decline in cases being performed by “chief residents” in general surgery, with overall major caseloads remaining unchanged. The authors interpreted the results as a possible redistribution of cases to younger resident levels (Christmas et al., 2009). A large scale review of national ACGME annual case log data also showed a decrease in operative volume in general surgery, although the numbers did not drop below the minimum requirement level (Simien et al., 2010). Furthermore, Simien and colleagues (2010) showed that case numbers in defined categories such as pancreatic, endocrine, and laparoscopic surgery actually increased. Besides the fact that laparoscopic procedures have increased over the past twenty years, this finding may also reflect that trainees select cases to attend depending on their educational importance. This assumption was supported by Hope et al.’s findings that case coverage by residents declined more markedly in some procedures such as dialysis access and soft tissue surgery than in others such as gastrointestinal or anorectal surgery following implementation of work hour directives (Hope et al., 2011). Furthermore, a study from the University of Mississippi Medical Centre General Surgery program showed that the case numbers as primary surgeon did not change, but the number of cases performed as a first assist dropped suggesting a redistribution of work load (Picarella, Simmons, Borman, Replogle, & Mitchell, 2011). Although case numbers as a primary surgeon may remain stable, the experience gathered through assisting appears to be reduced. This is concerning as, following principles of graded responsibility, trainees should ideal have an opportunity to first observe the procedure that is to be learnt before attempting steps themselves. The situation is naturally more aggravated in Europe, where stricter work-hour restrictions have been enforced. It is not surprising that a review of early implementation of a work-hour compliant schedule showed that in the U.K. 50 percent of trainees had missed operative opportunities due to the new schedule and that 85 percent attended procedures on their day off to compensate (Canter, 2011). Furthermore, Lonergan and colleagues (2010) reported that, in Ireland, less than 5 percent of basic surgical
trainees met their individual operative educational goals within a six-month rotation. Moreover, less than half of the trainees reached their targeted caseload numbers as assessed through mandatory reporting of operating logs (Lonergan, Mulsow, Tanner, Traynor, & Tierney, 2010). The problem with all estimations of caseloads, is that the studies rely on resident self-report data which may well result in an inaccurate representation of the topic as case numbers may be under-reported, or participation level could also be over-reported. For example, Snyder et al. (2012) showed that half of the residents responding to a survey on topics pertaining to teaching in the OR, admitted to logging procedures as a primary surgeon even when they felt they did not truly act as a primary surgeon during the operation (R. A. Snyder et al., 2012). This finding is concerning, as case logs still represent a method by which training programs continue to determine competence (R. A. Snyder, et al., 2012).

As repetitive learning encounters have become rarer in the OR, operative exposure may not culminate in surgical competence. An analysis of operative logs of graduating U.S. residents in 2003 identified that only 19 procedures had been performed at least ten times, with the most common cases completed being laparoscopic cholecystectomies (R. S. Chung, 2005). Similarly, Bell et al. (2009) showed that residents, graduating in 2005, only gained repetitive experience in 18 of the 121 procedures deemed important by program directors. These 18 procedures were performed at least ten times, but only nine different procedures were performed at least 20 times during the course of a five-year residency (Bell et al., 2009). Therefore, it seems logical that surgical education must focus on optimizing the learning benefit of each operative encounter, a notion that was recently emphasized by an international consensus statement on surgical education and training in light of the reduced work-hours which stated that every opportunity in the workplace needed to be used ensure the training of competent surgeons (Collins, 2011).

1.5.2 Societal Factors - Patient Safety

Surgeon educators must balance society’s need for new doctors against demands to ensure that the highest level of patient care is delivered to each and every patient (Raja & Levin, 2003). These contradicting requirements may appear unsolvable as in order to produce competent future
surgeons, trainees must be actively involved in patient care at all levels. Exposing patients to the trainees’ learning curve has led to concerns regarding patient safety issues, especially in the advent of laparoscopic surgery (Dent, 1991; Royston, Lansdown, & Brough, 1994; Tompkins, 1990). The introduction of simulation helped ease concerns as much of the learning curve can be mastered outside of the OR. Ultimately though, the OR cannot entirely be replaced making it inevitable that trainees gain experience through operating on real patients. Although the thought of becoming a “learning case” may be disconcerting, a recent survey of patients and their relatives revealed that the vast majority of respondents were not opposed to resident involvement in their case or care (Berg, Engel, Saba, & Hatton, 2011). Similar confidence in surgery performed by trainees was noted in the field of ophthalmologic procedures, with the vast majority of patients receiving surgery from a trainee stating that they were either very confident or completely confident in the trainee performing their case (Nehls, Ghoghawala, Hwang, & Azari, 2014). Although patients agreed with resident participation, the vast majority felt that the staff surgeon should be present throughout the procedure (Berg, et al., 2011). These opinions demonstrate that patients understand that training is necessary, but strongly feel that training should be performed under supervision. The need for supervision is also emphasized by regulating bodies such as the College of Physicians and Surgeons of Ontario (CPSO). Current regulations state that if trainees are to perform the majority of a case without direct supervision explicit patient consent is mandated (College of Physicians and Surgeons of Ontario, 2011). Interestingly, the survey by Berg et al. (2011) also revealed that only a minority of patients were aware of whether a resident would be involved in their care. This finding is surprising, as informing patients about resident involvement (even under supervision) could be considered a moral obligation (Raja & Levin, 2003).

A number of studies have been conducted to address the concerns regarding the influence of resident participation by evaluating patient outcomes (Advani, Ahad, Gonczy, Markwell, & Hassan, 2012; Castleberry et al., 2013; S. S. Davis, Jr. et al., 2013; Gorgun et al., 2014; Hernandez-Irizarry, Zendejas, Ali, Lohse, & Farley, 2012; Iordens, Klaassen, van Lieshout, Cleffken, & van der Harst, 2012; Rovito, Kreitz, Harrison, Miller, & Shimer, 2005; Seib et al., 2014; Tseng et al., 2011; Venkat, Valdivia, & Guerrero, 2014). The majority of these studies draw on the large American College of Surgeons National Surgical Quality Improvement Program database (ACS NSQIP). The database has been used to examine whether morbidity and
mortality rates and operative duration are influenced by resident involvement. Tseng et al. (2011) used the ACS NSQIP database entries from 2005 to 2007 to assess risk-adjusted morbidity and mortality rates in seven common general surgical procedure types. The authors found that resident participation resulted in an increase in morbidity by 14 percent with a 58 percent lower risk of perioperative mortality. Furthermore, resident participation was noted to prolong surgery by 16 to 31 minutes, although OR durations did not differ between postgraduate year (PGY) training level (Tseng, et al., 2011). The authors could not ascertain which mechanism led to the reduced mortality rate. A similar study drawing on the ACS NSQIP database entries between 2005 and 2008 examined the role of resident participation on outcomes in six different types of laparoscopic general surgical procedures (S. S. Davis, Jr., et al., 2013). This study also revealed an increased OR duration and morbidity rate, whilst mortality rate was reduced. As in the prior study, resident training level was not shown to correspond with OR duration (S. S. Davis, Jr., et al., 2013). In both studies, the magnitude of findings was not deemed clinically relevant by the authors. Two studies addressing the same question in adrenal surgery, both using the ACS NSQIP database, showed that resident participation either had no effect on outcomes (Venkat, et al., 2014), or led to a reduced risk of perioperative complications (Seib, et al., 2014), while OR duration increased in both studies. A further study looking at the same question in the field of laparoscopic inguinal hernia repair, used the ACS NSQIP database entries between 2007 and 2009 and also showed that OR duration increased, although in this study the increase in procedure duration correlated with increasing PGY training level. A possible explanation for this positive correlation given by the authors was that procedures with participation of lower training levels likely reflected an increased staff involvement. As no significant differences in postoperative morbidity or mortality were seen, the authors further questioned the relevance of the longer OR durations (Hernandez-Irizarry, et al., 2012). This inverse relationship between training level and outcome parameters was also shown by Gorgun et al. (2014) in laparoscopic colorectal surgery. These authors noted an increased length of hospital stay, sepsis rates and need for mechanical ventilation due to participation of fellows (PGY six and above)(Gorgun, et al., 2014). The subgroup analysis of PGY level revealed that complication rates were only increased in the fellow group, likely due to the higher degree of active participation whilst still having overall lower skill levels than staff (Gorgun, et al., 2014). This higher complication rate attributable to more senior trainees was also seen in advanced oncologic cases (Castleberry, et al., 2013). Like the previous studies, Castleberry et al. (2013) showed a decrease in mortality
corresponding to resident participation. Furthermore, the authors were able to demonstrate that resident participation resulted in a higher rescue rate of individuals that did suffer from a complication subsequently reducing mortality. The higher morbidity rates were attributable to higher rates of organ/-space infections (Castleberry, et al., 2013).

All these studies showed that although OR durations increased, complication rates either remained unchanged or increase without clinical relevance. The mortality was shown to be decreased in all studies likely due to resident involvement in postoperative care. Since all studies used the ACS NSQIP database, details on the actual amount of resident participation were not available. The studies could therefore only determine whether or not a resident was present, but not how much of the operation was completed by the trainees. In contrast two single-center studies conducted to assess whether resident involvement in advanced laparoscopic procedures (LRYGB) resulted in measurable increases in perioperative complications showed that active participation with a graded stepwise approach to learning the complex procedure did not lead to adverse patient outcomes (Iordens, et al., 2012; Rovito, et al., 2005).

1.5.3 Surgeons as Educators

Teaching expertise was long deemed to result from content knowledge. Over the last decades though, the understanding that dedicated teaching skills need to be developed, which are distinct from content expertise, has gradually evolved (Wilkerson & Irby, 1998). Despite the fact that surgeons play a central role as educators in the OR, surgeons rarely have any formal background in adult learning or teaching (Iwaszkiewicz, et al., 2008). Expert surgeons cannot simply be equated to good educators by the mere qualification attributed to the expertise in surgery. For example, surgical expertise is known to lead to automaticity in step execution, which may even in fact impede surgical education. A surgeon that completes operative steps in an automated fashion would not be aware of the need to explain individual substeps so that a trainee would be able to follow. This lack of explanation leads to an inadequate educational experience for the trainee (Butvidas, et al., 2011; Sullivan et al., 2008). Therefore, the surgeon as an educator needs
to become aware of his or her role as a teacher in the OR and modify behaviours accordingly. As described by Butvidas et al. (2011)

“teaching needs to become deliberate” (Butvidas, et al., 2011)[pg 388]

Thus, faculty development has been deemed essential to ensure that surgeons in the role of educators have the necessary skills and knowledge to deliver high-quality surgical education (Thomas, 2009). A number of different faculty development formats have been trialed and applied to improve the quality of surgical education. These include peer-coaching, workshop sessions, brief courses, seminars and dedicated fellowships (Irby & Wilkerson, 2003; Wilkerson & Irby, 1998). Although a number of dedicated fellowships and degree programs have been initiated to improve medical education (Fry, 2009), most faculty development initiatives are focused on brief training interventions as the interventions need to be integrated into the already busy work day of a clinician. For example, the RACS developed a three-hour course aimed at familiarizing surgical educators with concepts related to regular workplace-based assessments that represent an important component of the new surgical training program implemented in Australia in 2008 (Copertino, Blackham, & Hamdorf, 2010). The effectiveness of this workshop was assessed through participants’ reactions as measured in an online survey. This survey revealed that responding course participants felt that the intervention had better prepared them to conduct in-training assessments although the response rate was only 35 percent (Copertino, et al., 2010). Similarly, in an attempt to condense faculty development and help incorporate faculty training into clinical practice, Pernar and colleagues (2012) trialed the use of succinct teaching emails. The training intervention included weekly emails delivered over a nine-month period. The faculty members participating in this study though did not feel that the emails were useful as they felt that the content was too vague. Subsequently the study failed to show a significant improvement in teaching skills as reported in surveys of clinical clerks who had received clinical teaching from the faculty members (Pernar et al., 2012).

Surgeon educators need to receive feedback regarding their teaching skills so that they can improve when necessary. This feedback would need to be meaningful, from a credible source, and should be combined with instruction on how to achieve improvements (Centra, 1993). One method of informing surgeons of their teaching abilities may be through resident teaching
evaluations. Maker and colleagues (2004) designed a faculty evaluation form by asking the residents as the end-user to define the criteria of a surgical role model (Maker, Curtis, & Donnelly, 2004). This form was used by residents to assess faculty members’ teaching performance. The results of the resident assessments were shared with the faculty members in the form of specific feedback. Following the evaluation and feedback, the majority of faculty had improved and those that had failed to do so were taken off the teaching service (Maker, et al., 2004). The residents’ feedback had resulted in changed faculty behaviours including an increased rate of attendance at teaching activities, increased rate of delivering didactic teaching, and an increased rate of providing feedback to the residents (Maker, et al., 2004). A second study by the same group then investigated the value of ongoing faculty evaluations, again using resident evaluations (Maker, Lewis, & Donnelly, 2006). In this study, faculty members continued to improve throughout the evaluation periods with the individuals receiving the lowest initial evaluations improving the most (Maker, et al., 2006).

In addition to challenges associated with a lack of formal training in adult education, surgeons face conflicting interests arising from their multiple roles as educators, clinicians and researchers (Thibault, Neill, & Lowenstein, 2003). For example, surgeons must ensure the highest level of patient care whilst still enabling active trainee participation throughout the cases (Iwaszkiewicz, et al., 2008; Raja & Levin, 2003). The balance between patient care and training responsibilities may be difficult to achieve. In addition, surgical education requires investment of time. Time though may be limited as surgeons need to secure their own income through efficient work, as well as commit time to activities such as research that may be necessary to achieve academic promotion (Iwaszkiewicz, et al., 2008). Furthermore, since research funding has also become a means of augmenting salaries in academic surgery, surgeons may reduce the time invested for teaching to maximize the time available for research (Thibault, et al., 2003). A recent survey of educators in Australia and New Zealand revealed, that although the vast majority responding surgeons found their role in teaching rewarding, committing approximately 20 percent of their work time to educating trainees, significant barriers to teaching were identified (Collins, Smith, Lambert, & Hillis, 2011). The barriers described were similar to those detailed above and included a lack of dedicated (protected) teaching time due to conflicting work commitments, as well as a perceived lack of support from the institution or employers (Collins, et al., 2011).
Thus, it must be acknowledged that teaching in the OR comes at a cost (Babineau et al., 2004; Chamberlain, Patil, Minja, & Kordears, 2012; Sasor, Flores, Wooden, & Tholpady, 2013). The actual cost though remains difficult to assess and generalize. Although factors such as cost of OR time and facility charges have been used to quantify the estimated price of residency training, this approach may be inaccurate since many costs represent fixed amounts that would not be decreased even if residents were not involved in training (Babineau, et al., 2004). In contrast, the cost associated with loss of alternate income, experienced by a surgeon investing time to train residents, should be acknowledged as a relevant factor (Babineau, et al., 2004; Sasor, et al., 2013). This cost, termed “opportunity cost”, arises when a surgeon invests time in one activity (such as training a resident in the OR) and subsequently loses the chance to generate other sources of income (by booking more cases), or to undertake other activities such as research (Babineau, et al., 2004; Sasor, et al., 2013). This opportunity cost poses a threat to surgical education in the OR, as the educator must make a priority decision regarding conflicting interests.

1.5.4 Residents as Learners

Similar to the surgeon as an educator, the resident must fulfill in part conflicting roles. First, the resident as a trainee has an obligation to participate actively in learning activities in order to ensure adequate progression within the training program (Houston, Conn, & Rajan, 2011). Second, residents are required to deliver service to patients in their roles as physicians. As the majority of a resident's time is spent at the workplace, delineating what constitutes a learning activity and what is a pure service requirement may be difficult (Houston, et al., 2011; Sanfey et al., 2011). The difficulty in identifying learning experiences has also been recognized in the OR and has been postulated to contribute to the discrepancies in opinions of faculty and trainees regarding the amount of intraoperative learning experiences (Butvidas, et al., 2011). Surprisingly, discrepancies in views on the topic of educational versus service activities were also found between trainees and program directors (Sanfey, et al., 2011). A survey revealed that residents rated more activities as service requirements than the program directors did (Sanfey, et al., 2011). This divergent view on the balance between service and education, especially at the
level of program directors, is concerning as in competency-based programs goals and learning objectives should be clear to all parties to avoid disappointment and frustration.

Nevertheless, it is clear that residents will have to take on duties that do not have a clear educational goal. Thus, from a societal, economic view, residents represent a significant workforce that ensures daily patient care in most countries. Consequently, modifications in postgraduate education aimed at improving resident quality of life and education, such as granting more protected study time, or reducing work-hours will not only impact the residents, but will have repercussions within the work environment. As the active work force is reduced, other care providers will be required to compensate the loss of “manpower” (Houston, et al., 2011; Nuckols, Bhattacharya, Wolman, Ulmer, & Escarce, 2009; Nuckols & Escarce, 2005, 2012). These socio-economic conflicts of interest may influence how postgraduate medical education is structured and delivered, and may also contribute to the so called “hidden curriculum.” The hidden curriculum, as described by Frederic Hafferty (1998), includes all organizational cues that are not part of the defined curriculum that contribute to how values and priorities within the system are conveyed. Official and hidden curricula may suggest conflicting priorities (Hafferty, 1998). For example, although protected study time is anchored in many postgraduate programs, learners may opt not to attend educational sessions to prevent possible negative consequences due to not having provided unofficially expected service requirements whilst at the study sessions. Similarly, although learning in the OR is a recognized essential building block of surgical training, the environment in the OR may not be welcoming and hierarchical structures may prevent the learner from feeling empowered to ask questions and initiate learning (Monkhouse, 2010). Alarmingly, a survey found that residents would not actively do anything about negative learning experiences and were unlikely to confront their educator about the situation (Butvidas, et al., 2011). This inhibition needs to be acknowledged by surgeon educators as it may pose a threat to learning in the OR.
1.6 Summary

The surgical learning environment is complex. Trainees will spend the majority of their time in the workplace, acquiring skills through everyday experiences according to principles of experiential learning. Although modern training programs are structured to include formal curricular components, everyday learning in the OR bears many similarities to traditional apprenticeships. In addition, in current competency-based programs, trainees must demonstrate proficiency in the core competencies deemed of relevance for the surgical profession. This competency-orientated training mandates frequent assessments of skills and knowledge to monitor adequate progression and afford guidance throughout the training period. Competency assessments need to be granular, fair, objective, and feasible in the workplace context. These assessments are also essential since they form the basis of meaningful formative feedback which is an indispensable instructional tool in workplace-based learning.

Experiential learning in the OR has, historically, been used successfully to train surgeons. In the current era though, time spent in the OR has been reduced as a consequence of restructured work schedules, adherence to work-hour regulations, and additional curricular obligations outside the OR. As unguided discovery learning as a form of experiential learning may not be very effective, structuring discovery learning becomes paramount. Numerous instructional approaches have been shown to be effective in this environment, and should thus be incorporated into routine training in the daily OR. Analysis of the OR as a learning venue though, has shown that these instructional approaches are infrequently and inconsistently used in routine practice. Daily learning opportunities may thus be lost due to a lack of teaching structure.

In addition to a lack of teaching structure, several factors on an organizational, societal, and individual level may interfere with training in the OR. These factors, for example, include time constraints in hospitals which are compelled to ensure an economic functioning of the OR. Moreover, societal expectations that demand that each individual be treated with the highest level of care possible may appear contradictory to the obligation to train new surgeons. Further, the surgeon involved in intraoperative teaching may lack the knowledge or experience required of an educator, or may be influenced by conflicting interests such as personal development and financial securities. Lastly, the resident as a learner needs to be considered. Residents as adult
learners are assumed to be self-directed in their goal to acquire competence. Still it must be noted that many trainees may not seek out educational opportunities, nor critique poor educational experiences. Further, physicians lack the ability to self-assess performance and, therefore may not be capable of determining learning needs. As a result, external feedback and guidance remains essential in ensuring all relevant skills are acquired during training.
2
PROBLEM STATEMENT, RESEARCH OBJECTIVES, AND HYPOTHESES

Within this chapter, the knowledge gap is described and the research problem defined. The aims, objectives and hypotheses of the project are stated.
2.1 Problem Statement/Rational

Structured surgical training is vital to ensure that the next generation of surgeons is equipped with the skills necessary to guarantee safe patient care, as well as the skills required to ensure effective ongoing professional development. The core content of what surgeons are to be taught is anchored in training frameworks regulated by local medical authorities. Still, details of how this content is to be taught, including which instructional strategies are to be used, remain at the discretion of the individual programs and local educators. Even though the vast majority of modern training programs are focused around competency-based, structured training and include defined curricular activities, elements of apprenticeship learning still remain dominant within workplace-based education. In this environment, surgical trainees gain knowledge and experience through hands-on practice under direct supervision and guidance of qualified surgeons. Learning within this type of relationship though requires time. This classical approach may no longer guarantee skill development due to increasing time constraints within the OR. These time constraints are the consequence of competing economic interests, reduced resident work-hours, and limited resident OR time as a result of other duties or curricular activities. In addition, the quality of training within the OR depends heavily on the teaching skills of the individual supervising surgeons. Since not every surgeon is a natural teacher or has received dedicated training to prepare for the role of a surgeon educator, educational quality will remain variable. Teaching behaviours such as effective communication using instructing and questioning skills, preoperative educational briefing, postoperative case debriefing, feedback, and coaching remain inconsistent. This lack of consistency contributes to the perception that the OR as a learning venue is an unstructured discovery learning environment. Unguided discovery learning though has been shown to be ineffective. Therefore, in an era of reduced opportunities to participate in operative procedures, instructional approaches must be deliberately applied to ensure that each and every procedure performed in a training environment serves the purpose of educating the future surgeon. Especially at an advanced stage of skill acquisition, beyond the phase of learning basic skills and surgical steps, the training surgeon will require focused teaching within the OR capitalizing on each procedure as a learning experience. New strategies will be necessary that target individual learning needs. Personalized approaches have a longstanding tradition in other fields, such as sports training, as a means to maximize an
individual's potential. Surgeons may equally be viewed as high-performance individuals that should continuously strive for excellence. Thus, concepts such as coaching may represent a practical methodology to structure experiential learning in the OR, integrating assessment, planning, goal setting, debriefing and feedback as highly effective teaching behaviours.
2.2 Research Aims and Objectives

The aim of this project was to design a technical training intervention for surgical residents and fellows, applicable to laparoscopic training in the OR. This training intervention was intended to be structured around coaching frameworks and based on performance analysis, individualized feedback, debriefing and behaviour modeling. The goal of training was to improve laparoscopic technical skill in the clinical environment.

Gaps in knowledge and lack of methodologies had to be addressed to accomplish this aim. These gaps included the absence of an objective assessment method for identifying technical errors in laparoscopic surgery, lack of frameworks and definitions guiding acceptable surgical task execution in laparoscopic surgery, and lack of an evidence-base for teaching examples that could be used in didactic sessions.

Thus, five distinct objectives were addressed to accomplish the aim of designing a meaningful new training intervention:

Objective 1: To develop an assessment tool to enable reliable measurement of technical errors in laparoscopic surgery, specifically LRYGB.

Objective 2: To establish clinically relevant definitions of common technical errors in laparoscopic surgery.

Objective 3: To characterize common intraoperative error-event mechanisms representing “near misses” in routine complex laparoscopic procedures.

Objective 4: To assess feasibility of regular surgical performance analysis using global, procedure-specific and technical error analysis in order to tailor individualized training in the form of surgical coaching.

Objective 5: To assess the transfer of skills of “coached” and “conventionally” trained residents in the real OR.
2.3 Hypotheses

2.3.1 Null Hypothesis (H₀)

I. H₀: There will be no difference between surgical technical skill as measured by global rating scales between trainees receiving coaching and those undergoing conventional training during a two month minimally invasive surgery (MIS) rotation.

II. H₀: There will be no difference in the frequency of observed technical errors between trainees receiving coaching and trainees undergoing conventional training during a two month MIS rotation.

2.3.2 Alternate Hypothesis (Hₐ)

I. Hₐ: Surgical coaching will improve participants' technical skills as measured by global rating scales

   Hₐ: Surgical coaching will decrease participants' technical skills as measured by global rating scales

II. Hₐ: Surgical coaching will improve participants' technical skills as measured by decreased technical error counts

   Hₐ: Surgical coaching will decrease participants' technical skills as measured by increased technical error counts
A prerequisite for meaningful performance feedback is the ability to comprehensively and objectively assess performance. In technical skills assessment, several aspects of technical performance can be evaluated, such as global and procedure-specific skill as well as technical task execution through error analysis. This chapter details the development of a framework for technical error analysis in laparoscopic surgery. Validity evidence for the new tool is gathered in the context of LRYGB procedures.

The contents of this chapter have been published in:

*Error rating tool to identify and analyse technical errors and events in laparoscopic surgery*

Bonrath EM, Zevin B, Dedy NJ, and Grantcharov TP


© 2013 British Journal of Surgery Society Ltd. Published by John Wiley &Sons Ltd.

The use of the manuscript is by permission of the copyright holder.
3.1 Abstract

**Background:** Surgical error analysis is essential for investigating mechanisms of errors, events and adverse outcomes. Furthermore, it provides valuable information for formative feedback and quality control. The aim of the present study was to design and validate a technical error rating tool in laparoscopic surgery.

**Methods:** The framework consisted of nine task groups and four error modes. Unedited videos of laparoscopic Roux-en-Y Gastric Bypass procedures were rated and analyzed. The Objective Structured Assessment of Technical Skill (OSATS) global rating scale was used to assess technical skill. The incidence of errors and of injuries (events) were the main outcome measures, and were used to calculate the reliability, and construct and concurrent validity of the instrument.

**Results:** Two observers analyzed 25 procedures. Inter-rater reliability was high regarding total number of errors (intraclass correlation coefficient (ICC) 0.90) and events (ICC 0.85). The median (interquartile range) error rate was 35 (26-44) and the event rate 3 (2-3) per procedure. Error frequencies and OSATS scores correlated significantly in all operative steps ($r_s = -0.75$ to $-0.40$, $P = < 0.001$-0.046). Surgeons demonstrating high OSATS scores had lower median (i.q.r) error rates than surgeons with low scores in three of four steps: measuring bowel (4 (2-7) *versus* 10 (9-11); $P = 0.004$), jejunojejunostomy formation (5 (2-6) *versus* 10 (9-11); $P = 0.001$) and pouch formation (4 (3-6) *versus* 9 (5-12); $P = 0.004$).

**Conclusions:** The proposed error rating tool allows an objective and reliable assessment of operative performance in laparoscopic gastric bypass procedures.
3.2 Introduction

Surgery is a high-performance, high-risk operation, and human error and adverse events have been a subject of discussion and research for many years. Early reports were published in the 1960s and led the way towards a culture of error analysis (Beaty & Petersdorf, 1966; Kohn, et al., 2000; Osborn, 1967; J. T. Reason, 1990). Analysis in a surgical setting revealed significant errors in surgical treatment and perioperative management. This highlighted the importance of recognizing and managing surgical error in order to improve patient safety (Couch, Tilney, Rayner, & Moore, 1981).

For many years, error analysis was based predominantly on retrospective chart reviews of closed malpractice cases (A. A. Gawande, et al., 1999; S. O. Rogers, Jr., et al., 2006). The data showed that more than half of documented adverse events were considered potentially avoidable (A. A. Gawande, et al., 1999). These types of study enabled a root-cause analysis approach to surgical risk management. However, owing to the retrospective nature of the data analysis, near miss situations were neither acknowledged nor investigated. Prospective analysis of technical surgical errors demonstrated that even minor intraoperative events can lead to a higher rate of patient morbidity and mortality (de Leval, et al., 2000). It has been advocated that surgeons should analyse their own errors in order to learn from them (Rebasa, et al., 2009). Despite the need for surgical error identification and analysis, there is a lack of methods to facilitate assessment. This is largely a result of variability and complexity in methodologies, as well as limitations regarding objectivity, validity, and reliability (Bonrath, Dedy, Zevin, & Grantcharov, 2013a).

A valid, reliable and clinically applicable error assessment tool is a prerequisite for the implementation of technical error analysis into clinical practice and surgical education. The aim of the present study was to design and validate a generic error rating tool (GERT) for objective assessment of technical errors in laparoscopic surgery.
3.3 Methods

Research ethics approval for this study was obtained from the local research ethics board. For the purposes of the investigation, the following definitions of the terms “error” and “event” were chosen to facilitate standardization of what observations should be included in the analysis. In this study the error represents the mechanism of unintended or deviated technical task execution, and the event the resulting injury or damage. Definitions of error taken into consideration were:

“an error is the failure of planned actions to achieve their desired goal” (J. Reason, 1995)[p 81],

“an act, assertion, or belief that unintentionally deviates from what is correct, right, or true” (Editors of the American Heritage Dictionaries, 2011),

and “disrespect of basic surgical principles”.

The definition of event was derived from World Health Organization guidelines:

“Any deviation from usual medical care that causes an injury to the patient or poses a risk of harm” (World Health Organization, 2005)[p 9]

In addition, the definition of an action that may require additional measures to avoid an adverse outcome (Barach et al., 2008), and derived from the Institute of Medicine's report on human error (Kohn, et al., 2000), was used.

Design of Generic Error Rating Tool

Various approaches to describe errors in surgery and other high-risk industries were critically reviewed and evaluated (Bonrath, Dedy, et al., 2013a). Following literature analysis, a proposed framework (Embrey, 1986), modified for use in the medical field (Joice, et al., 1998), was selected as the basis for the GERT. Four modified error modes were defined: inadequate use of
force or distance (too much); inadequate use of force or distance (too little); inadequate visualization; and wrong orientation of instrument or dissection plane. Each error mode could potentially be encountered in a variety of different surgical tasks. Nine generic surgical tasks were defined, applicable to laparoscopic surgery, during which an error could be committed: abdominal access; use of retractors; use of energy devices; grasping and dissection; cutting, transection and stapling; clipping; suturing; use of suction; and other unclassified. The latter group was included to allow for any additional remarks. The nine surgical task groups were combined with the four error modes in a checklist format (Appendix 1).

In the present study, the concept of an error-event hierarchy was followed. In this hierarchy, an error represents the lowest level of how a task could be executed incorrectly, whereas an adverse outcome represents the highest level (Figure 4). Only the two lowest levels, error modes and technical events, were the focus in the present study.

**Figure 4: Error-event-outcome hierarchy, with examples.**

[MOF] multiple organ failure
Determination of errors and events

LRYGB was selected to determine the reliability and validity of the GERT as this represents a complex procedure for which the surgeon must perform multiple different tasks for successful completion. Two observers with expertise in surgical education and bariatric surgery independently reviewed unedited videos of LRYGB, and tabulated observed errors and events. Videos were retrieved arbitrarily from an educational video library containing anonymized recordings of 250 LRYGB procedures. The observers were instructed regarding the definitions of error modes and events by reviewing video examples. Observed errors and events were time-stamped to the moment of occurrence. The total number of errors observed during LRYGB was the main outcome measure for this study. Events related to technical errors were recorded and analyzed to determine causal patterns.

Inter-rater reliability

The total number of errors and events recorded by each observer was used to calculate the inter-rater reliability of the GERT. The calculation was then repeated for each generic surgical task category.

Construct and concurrent validity

To validate the GERT, objective definitions of surgical skill and expertise were required. Owing to anonymous nature of the recordings, experience of the primary surgeon (such as consultant or trainee) was not known to the observers or researchers. As a result, a score on the OSATS global rating scale (Martin, et al., 1997) was used to group surgeons into high-performer and low-performer groups. A high performer was defined as an individual who achieved an OSATS score of at least 28 (80 percent of the maximum score). A low performer was defined as an individual with an OSATS score of less than 28. This cut-off was chosen based on the results of previous studies that reported consultant surgeons scoring at least 80 percent on global OSATS-rating scales (Beard, et al., 2007; Curry, et al., 2012; Kumar, et al., 2012).
To account for the different surgeons completing different steps of the operation, objective assessment of technical ability of the primary surgeon was conducted for the following four steps of the operation: measuring bowel (MB), jejunojejunostomy formation (JJ), gastric pouch creation (pouch) and GJ formation (Table 5). The lowest observed skill level during a step was used to determine the OSATS score for that step.

Differences in total number of errors between high- and low-performance surgeons for each step were used to assess the construct validity of the GERT. The total score on the OSATS scale and the total number of errors on the GERT were correlated for each operative step to determine the concurrent validity of the GERT.

<table>
<thead>
<tr>
<th>Start (S) and endpoint (E)</th>
<th>Standard skills and tasks observed</th>
</tr>
</thead>
</table>
| Measuring bowel | S: Identification of ligament of Treitz  
| E: Cutting of stay suture | Grasping: measuring biliopancreatic and Roux limbs, handling of bowel  
| | Stapling: use of linear stapler  
| | Clipping: marking biliopancreatic limb  
| | Suturing: placement of stay stitch  
| Jejunojejunostomy formation | S: Cutting of stay suture  
| E: Cutting of sutures after primary enterotomy defect closure | Use of energy device: enterotomy creation  
| Corrective action: placement of additional reinforcement sutures after primary closure following inspection of anastomosis | Stapling: use of linear stapler  
| | Suturing: closure of enterotomy defect  
| Gastric pouch creation | S: Introduction of energy source for dissection of lesser curve  
| E: Confirmation of complete transection of stomach | Grasping: handling of stomach  
| | Stapling: use of linear stapler  
| | Use of energy device: dissection of lesser curve and phreno-esophageal ligament  
| | Dissection: development of posterior plane for pouch creation  
| Gastrojejunostomy formation | S: Application of energy source for creation of enterotomy in Roux limb  
| E: Cutting of sutures after primary closure of enterotomy defect | Use of energy device: enterotomy creation  
| Corrective action: recorded if additional reinforcement sutures are placed after primary closure during inspection of anastomosis | Stapling: use of linear stapler  
| | Suturing: closure of enterotomy defect  
| Other | All time periods outside defined steps  
| | Abdominal access: trocar insertion and removal  
| | Use of retractors: insertion of liver retractor  
| | Clipping: haemostasis  
| | Suction: as required during haemostasis, or inspection of anastomosis or pouch results  

Table 5: Operative steps of laparoscopic Roux-en-Y gastric bypass
Statistical analysis

Continuous variables are reported as median (i.q.r.) unless indicated otherwise, and differences between groups were evaluated by the Mann-Whitney $U$ test. Differences in categorical data between groups were analysed by Fisher’s exact test. Inter-rater reliability was calculated by means of the ICC, and correlations were calculated by Spearman correlation. $P < 0.050$ was considered significant.

3.4 Results

Error and event rates

Video recordings of 25 LRYGB procedures were analysed. A mean total of 915 errors between two observers (observer 1, 926 errors; observer 2, 904 errors) and 63 events were recorded (Table 6). A median of 35 (26-44) errors and 3 (2-3) events per procedures was calculated from the averaged scores of the two observers. Errors were observed most frequently during the tasks “suturing” (44.5 percent) and “grasping and dissection” (40.9 percent) (Table 6, Figure 5).
<table>
<thead>
<tr>
<th>Procedure</th>
<th>Observer 1</th>
<th>Observer 2</th>
<th>Intra-class correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal access</td>
<td>45</td>
<td>45</td>
<td>0.85</td>
</tr>
<tr>
<td>Use of retractors</td>
<td>2</td>
<td>3</td>
<td>0.39</td>
</tr>
<tr>
<td>Use of energy devices</td>
<td>20</td>
<td>29</td>
<td>0.51</td>
</tr>
<tr>
<td>Grasping and dissection</td>
<td>390</td>
<td>359</td>
<td>0.91</td>
</tr>
<tr>
<td>Cutting, transection and stapling</td>
<td>16</td>
<td>29</td>
<td>0.89</td>
</tr>
<tr>
<td>Clipping</td>
<td>26</td>
<td>22</td>
<td>0.80</td>
</tr>
<tr>
<td>Suturing</td>
<td>410</td>
<td>405</td>
<td>0.82</td>
</tr>
<tr>
<td>Use of suction</td>
<td>6</td>
<td>1</td>
<td>n.a.</td>
</tr>
<tr>
<td>Other unclassified</td>
<td>11</td>
<td>11</td>
<td>0.59</td>
</tr>
<tr>
<td>Total no. of errors observed in 25 procedures</td>
<td>926</td>
<td>904</td>
<td>0.90</td>
</tr>
<tr>
<td>Total no. of events observed in 25 procedures</td>
<td>63</td>
<td>63</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 6: Number of errors and events in 25 procedures.

[n.a.] Not applicable
Figure 5: Errors per procedure noted by two observers.

Box plots show median (bold horizontal line within box), interquartile range (box) and 10-90\textsuperscript{th} percentile (error bars). Circles and asterisk represent outliers (more than 1.5 but less than 3 interquartile ranges from end of box) an extreme outliers (more than 3 interquartile ranges from end of box) respectively.

The most frequently observed error mode was “too little force or distance” during the surgical task “grasping and dissection” (observer 1, 332 errors; observer 2, 297 errors). The distribution pattern of the individual error modes within the two most frequent surgical task groups (“suturing” and “grasping and dissection”) are shown in (Figure 6).
Figure 6: Distribution of error modes for each observer.

Results are shown for analysis of all 25 procedures for error task groups a “grasping and dissection” (intraclass coefficient (ICC) 0.91), and b “suturing” (ICC 0.82).

Inter-rater reliability

Inter-rater reliability was very high for total number of errors (ICC 0.90) and events (ICC 0.85). Similarly, high agreement was observed for the total number of errors during the surgical tasks “abdominal access”, “grasping and dissection”, “cutting, transection and stapling”, “clipping” and “suturing” (all ICC more than 0.80). Moderate agreement existed for the number of errors observed during “use of energy devices” and the group “other unclassified” (ICC 0.50-0.60). Low agreement was found for total number of errors observed during “use of retractors” (ICC 0.39). Statistical assessment of inter-rater reliability for “use of suction” was not possible owing to the infrequent observation of errors during this task (Table 6).
Construct and concurrent validity

For the operative steps MB, JJ and pouch creation, a difference in the median number of errors committed by high- and low- performance surgeons was noted, with low-performance surgeons making a greater number of errors (Table 7). Significant correlations between the OSATS scores and total number of errors on the GERT tool were demonstrated for all operative steps (MB: $r_S = -0.46, P = 0.023$; JJ: $r_S = -0.75, P < 0.001$; pouch $r_S = -0.40, P = 0.046$; GJ: $r_S = -0.47, P = 0.018$).

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Error frequency$^*$</th>
<th>P$^\dagger$</th>
<th>No. of procedures requiring anastomosis corrective action</th>
<th>Duration of corrective action [h]$^\ddagger$</th>
<th>Total duration of operative step [h]$^*$</th>
<th>P$^\ddagger$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measuring bowel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low performer (n = 6)</td>
<td>10 (9–11)</td>
<td>0.004</td>
<td>n.a.</td>
<td>n.a.</td>
<td>968 (805–1163)</td>
<td>0.092</td>
</tr>
<tr>
<td>High performer (n = 18)</td>
<td>4 (2–7)</td>
<td></td>
<td></td>
<td></td>
<td>734 (481–1032)</td>
<td></td>
</tr>
<tr>
<td>Jejunojejunostomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low performer (n = 11)</td>
<td>10 (9–11)</td>
<td>0.001</td>
<td>3</td>
<td>248 (243–305)</td>
<td>1330 (960–1618)</td>
<td>0.107</td>
</tr>
<tr>
<td>High performer (n = 14)</td>
<td>5 (2–6)</td>
<td></td>
<td>3</td>
<td>197 (112–408)</td>
<td>1029 (615–1366)</td>
<td></td>
</tr>
<tr>
<td>Pouch formation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low performer (n = 5)</td>
<td>9 (5–12)</td>
<td>0.004</td>
<td>n.a.</td>
<td>n.a.</td>
<td>1028 (953–1497)</td>
<td>0.371</td>
</tr>
<tr>
<td>High performer (n = 20)</td>
<td>4 (3–6)</td>
<td></td>
<td></td>
<td></td>
<td>907 (780–1224)</td>
<td></td>
</tr>
<tr>
<td>Gastrojejunostomy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low performer (n = 6)</td>
<td>9 (6–12)</td>
<td>0.406</td>
<td>5</td>
<td>416 (86–1296)</td>
<td>1725 (1439–2323)</td>
<td>0.262</td>
</tr>
<tr>
<td>High performer (n = 17)</td>
<td>7 (5–11)</td>
<td></td>
<td>7</td>
<td>285 (120–619)</td>
<td>1575 (1301–1810)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Error frequency in relation to surgeon’s skill level as determined by the Objective Structured Assessment of Technical Skill global rating scale

[*] Values are median (interquartile range) and [†] median (range). Low performers had Objective Structured Assessment of Technical Skill (OSATS) score below 28 and high performers had an OSATS score of 28 or more. [n.a.] not applicable. [‡] Mann-Whitney U test and [$\S$] Fisher’s exact test.

Procedure duration and error rate

There were no differences in time required to complete a specific step of the procedure between the high- and low- performance surgeons (Table 7). The longest duration of time was spent on GJ (high-performance group: 1575 (1301-1810) s; low-performance group: 1725 (1439-2323) s), and the shortest on MB (734 (481-1032) and 968 (805-1163) s respectively) (Table 7). For all
four procedural steps significant correlations between time to complete the step and number of errors were noted (MB: $r_S = 0.58$, $P = 0.002$; JJ: $r_S = 0.59$, $P = 0.002$; pouch: $r_S = 0.41$, $P = 0.040$; GJ: $r_S = 0.44$, $P = 0.027$).

Error-event patterns

Each observer noted a total of 63 events in the 25 procedures (ICC 0.85). Comparison of the descriptions and timing of these events revealed that the observers identified and described identical events in 35 instances (56 percent). Observers 1 and 2 each noted additional 28 events that were different in terms of time point of observation or type of error assumed to underlie the event. Therefore, 91 distinct events were analysed. The most frequent events observed were bleeding from staple lines, mesentery, or omentum (28 events, 31 percent), serosal tears (17, 19 percent), burning of non-target structures (13, 14 percent) and ripped sutures (11, 12 percent). The most common underlying mechanisms for these events were “too much force or distance” when suturing (21, 23 percent), “too much force or distance” when grasping (21, 23 percent) and “too little force or distance” when grasping (9, 10 percent). The majority of events resulted from errors in the basic tasks of suturing (33, 36 percent) and grasping (32, 35 percent). All observed events were noted by the surgical team and rectified during the initial surgery.
3.5 Discussion

The present study has demonstrated the inter-rater reliability and construct and concurrent validity of a tool for quantitative and qualitative assessment of technical errors in laparoscopic surgery. The tool allowed the identification of the root causes of adverse events by analysing error-event patterns.

Multiple authors have used error analysis as a quantifiable objective parameter to assess surgical proficiency and skill (Ahlberg, et al., 2007; Hwang, et al., 2006; Pugh, et al., 2010; Seymour, et al., 2002). Error analysis has been applied extensively in VR training, with most simulators programmed to report such metrics. However, they have limited value in clinical training as a source for formative feedback (Stefanidis, et al., 2009). Although the need for additional parameters to facilitate comprehensive skills assessment has been recognized, current approaches lack generalizability (Bonrath, Dedy, et al., 2013a).

The initial error analysis approach (Joice, et al., 1998), utilizes generic error modes, but is based on a predefined order of operative steps within a procedure. This complex methodology, defined as hierarchical task analysis, has been used to evaluate a few procedures and mainly in a research setting (Talebpour, et al., 2009; Tang, Hanna, Bax, et al., 2004; Tang, Hanna, Joice, et al., 2004). Application of this GERT, does not require prior procedure deconstruction into operative steps, which increases its feasibility and generalizability.

The GERT demonstrated high inter-rater reliability, and can be assumed to deliver consistent measurements of total error and event frequencies when used by different raters. Inter-rater agreement was high in five of the eight task groups. The inter-rater reliability was low in the category “use of retractors”, possibly attributable to the very low frequency of observations. The moderate agreement in the task group “use of energy devices” could have been due to the overall lower frequency of these errors. Furthermore, delineation of erroneous use of an energy device and normal variation in task execution might have proven difficult for the observers and require more detailed definitions. Therefore, both low incidence and discreetness probably led to the lower level of agreement for this error task group.
Previous authors have focused predominantly on error analysis in the context of basic procedures such as the laparoscopic cholecystectomy (Bonrath, Dedy, et al., 2013a). On average, expert surgeons committed two to 11 errors, and novices between five and 18 errors, per procedure (Mishra, et al., 2008; Seymour, et al., 2004; Tang, et al., 2005; Tang, Hanna, Joice, et al., 2004).

In the present study, two to three times as many errors were identified compared with previous reports. This could possibly be related to the longer duration and complexity of LRYGB compared with laparoscopic cholecystectomy. The present analysis demonstrated notable variability in total error scores reflecting performance differences within standardized procedures.

Although some of the observed errors are unlikely to cause patient harm, they were included to adhere to the definitions, and maintain consistency and objectivity. Analysis of error and event patterns demonstrated that even minor errors, such as slipping off the bowel with an atraumatic grasper, could cause an injury that required repair.

Identical injuries and mechanisms were described by the observers in only just over half of the recorded events. This may have been a result of observer reaction time lag, identification of the same events at differing times within the procedure, or differences in what was presumed to be the underlying mechanism. Regardless of potential redundant event scoring, this information can be used for formative feedback, as recognizing error-event patterns represents an important teaching and safety concept that highlights the relevance and clinical consequence of different, seemingly minor, intraoperative mishaps.

In the present study OSATS scores rather than experience or rank was chosen to calculate concurrent validity. Concurrent validity has been defined as a measure of relationship between a given test score and a score obtained by another valid scale measuring the same underlying construct. The OSATS was chosen as a valid measure of technical skill and surgical technical competence. Error frequency, detected by the GERT was hypothesised to reflect this same construct. The results of the present study, confirmed a significant relationship between the surgeon's technical skills score and the number of errors observed in all operative steps of LRYGB supporting this theory. For construct validity, a cut-off level on the OSATS global rating score was defined to assign participants to high- and low-performer groups. Using this cut-off, the study demonstrated construct validity for three of the four operative steps.
Time has frequently been used as a performance metric during simulation training and assessment (Fraser et al., 2003; Stefanidis, et al., 2009). The results of the present study showed a correlation between number of errors and length of procedure, but no correlation between operative performance score as measured on the OSATS scale and length of procedure. The results indicated that the skill level itself in an educational environment did not lead to significant or clinically relevant increases in procedure duration, and that time should not be the primary outcome when assessing procedural proficiency.

The present study had a number of limitations. Although the aim was to design a generic tool, only one procedure was used to validate the approach formally. A complex laparoscopic procedure was chosen with the expectation errors and events would be observed during multiple surgical manoeuvres. As this study investigated only the validity of the tool, no patient outcome data were collected. Correlation of surgical errors with measures such as training status, surgeon volume, and ultimately, patient outcome can be addressed only once a valid method of analysis is available for use in research and education. Furthermore, and of importance, the GERT is a tool for assessment of technical error; therefore, errors outside the operative field (equipment failures, non-technical skills, patient factors) were not part of the study and subsequently not assessed.

In spite of these shortcomings, the GERT represents an objective, reliable and easy-to-use tool for assessment of surgical errors and surgical performance. It can be used to provide meaningful feedback to both novice and experienced surgeons, and may contribute to early recognition of errors and event patterns that may lead to adverse outcomes.
This chapter describes the Delphi consensus process used to obtain an overview of definitions and examples of technical errors in laparoscopic surgery relevant to surgical education. These definitions are required to standardize assessment and training focusing on error analysis. Furthermore, exploring expert opinion is essential in order to identify divergent views and controversial topics that may impact generalizability of teaching content.

The contents of this chapter have been published in:

*International consensus on safe techniques and error definitions in laparoscopic surgery*

Bonrath EM, Dedy NJ, Zevin B, and Grantcharov TP

Surgical Endoscopy 2014 May; 28(5):1535-1544


© Springer Science+Business Media New York 2013

With kind permission from Springer Science+Business Media

4.1 Abstract

*Introduction:* Definitions of errors and poor technique in laparoscopic surgery are lacking in modern clinical practice. As a result, educators often base their teaching on personal experience and individual preferences. The objective of this study was to achieve expert consensus regarding these definitions in order to provide a framework for a standardized approach to teaching safe technique and avoiding common errors in laparoscopic surgery.

*Methods:* A Delphi survey was conducted with an international panel of experts in laparoscopic surgery. Survey items for definitions and examples of errors and resulting injuries (events) were derived from literature reviews and procedural observations. An online platform was used to administer the survey. Experts were requested to rate their level of agreement regarding survey items on a 5-point Likert-type scale; additional comments were facilitated through free-text entries. Consensus was defined as Cronbach's $\alpha > 0.70$. Items that were rated $\geq 3$ (“somewhat agree”) by 75 percent or more of the panel rated were included in the consensus list. The Delphi-process was continued until all subsections of the survey met the defined consensus level.

*Results:* Two survey rounds were completed with 33 experts from 12 countries (round one) and 25 experts from 9 countries (round two). Overall consensus was high for both rounds ($\alpha = 0.9$). Seventeen definitions and 39 examples of errors and events were included in the final consensus list.

*Conclusions:* Standardized definitions and examples of technical errors in laparoscopic surgery were established in a consensus-based approach. These definitions can serve as a uniform nomenclature, and can be used by educators as a reference guide to ensure standardization in surgical training and performance assessment.
4.2 Introduction

Modern surgical training is still deeply rooted in last century's traditional Halstedian apprenticeship model. Although educational research has had great achievements since those early days, many educational concepts in open surgery remain unchanged, including the teaching of basic surgical principles of tissue handling and operative technique. The introduction of laparoscopic surgery in the late 1980s, led to a shift in traditional teaching concepts as general surgeons were faced with the need to learn a unique skill set in a method that gained rapid recognition and wide implementation without access to extensive training and methods for quality assessment.

During the implementation era of laparoscopic surgery, concerns were voiced regarding a lack of teaching approaches (Dent, 1991; Royston, et al., 1994; Tompkins, 1990). Consequently, training and accreditation requirements were devised and simulation training, observerships, and structured courses were endorsed (Dent, 1991; Royston, et al., 1994; EAES, 1994; SAGES, 1991). However, learning the novel skill from an experienced mentor represented a vital training requirement then and continues to play a significant role in current surgical training in numerous surgical disciplines, raising questions about the basis for credentialing of experts and mentors (Abboudi et al., 2011; Entezami, Franzblau, & Chung, 2012; Kron, 2011). In fact, a recent analysis of the literature on mentor-mentee relationships in surgery revealed a lack of qualified mentors due to inadequate mentorship training (Entezami, et al., 2012). Similarly, concerns must be voiced regarding the quality of the content of training courses, particularly in view of the vast number of postgraduate training courses commercially offered to surgeons to acquire new laparoscopic skills. Some standardization of training content is required to ensure acquisition of safe techniques that are universally accepted within the laparoscopic community. However, the question arises as to how a minimum level of standardization can be assured if debates on correct and incorrect techniques are still ongoing and definitions of errors in laparoscopic surgery are lacking (Bonrath, Dedy, et al., 2013a; Cuschieri, 2003).

The objective of this study was to establish a widely accepted and clinically relevant set of definitions of common technical errors in laparoscopic surgery that would serve as a reference guide for educators and practitioners.
4.3 Methods

*Ethics*

Approval for this study was obtained from the research ethics board of the University of Toronto.

*Study design: Delphi methodology*

A modified Delphi methodology was selected to attain consensus through expert opinion on the definitions of surgical technical error. In the modification used in this study, a semistructured first survey round combined "closed-ended" items with options for free-text entries. This modification represents an established methodological variation of the Delphi technique (Duffield, 1993; Hsu & Sandford, 2007; Palter, MacRae, & Grantcharov, 2011). During the first survey round, the structured items were subject to expert ranking of perceived level of relevance. Statistical evaluation of these items was conducted and provided to panelists as a form of clarification and feedback during the second survey round. This process of feedback represents an essential methodological component of the Delphi process and was performed to allow the individual panelist to reassess his or her primary rating in the context of the group's opinion. This process was repeated until a predefined level of agreement among panel members was achieved and consensus was obtained for the overall survey as well as for all major subsections. The free-text entries submitted by the panelists during the first round were subjected to qualitative analysis. The responses were assessed for emerging major themes as well as for topics that had not been fully reflected in the structured items of the first round. These were subsequently used to generate additional items for the second survey round. Prior to each round of the survey, items were piloted within the research team (three fellows, two residents and one staff surgeon) to assess for clarity and ambiguity.
Survey administration and item generation

The structured items used in the first round of the survey were based on the results of a literature review on the topic of surgical technical error (Bonrath, Dedy, et al., 2013a). Furthermore, error examples obtained from procedural observations of advanced laparoscopic procedures were also included. The survey was created using an online platform (SurveyMonkey®) with a link to the website sent via e-mail to the identified experts through their publicly available email addresses. In the recruitment email, the purpose of the study was explained, and interested participants were requested to follow the link to the first survey round. All initial nonrespondents were requested three separate times by email to participate in the study. Only those participants who completed the first round of the survey were contacted to complete the second round.

Participants

Content experts in the field of advanced laparoscopic surgery were recruited as panelists for this study. These individuals were identified through their public affiliations with national and international societies for surgery, laparoscopy and obesity surgery, as well as through their publications in peer-reviewed journals on the topic of advanced laparoscopy, surgical education or error analysis. Having an active laparoscopic practice and involvement in clinical training of surgical residents and fellows were prerequisites for selection.

Ranking and determining consensus

The expert panel was requested to rate their level of agreement with statements regarding surgical errors on a 5-point Likert-type scale. To avoid a neutral ranking point, the scale was constructed as a right-packed interval continuum with the values defined as follows: 1 = “disagree”, 3 = “somewhat agree”, 5 = “fully agree”. In the free-text option, experts were encouraged to add error definitions based on their clinical experience. Responses were confidential and anonymous to other participants. Members of the panel were unaware of the identity of other panel members.
Consensus was predefined as internal consistency within the opinions of the panel as assessed by the individual ratings on the Likert-type scale. This approach has been used in previous Delphi surveys and is based on work supporting the role of Cronbach’s α as a measure for consistency in the context of Delphi studies by Graham et al. (2003) (Graham, Regehr, & Wright, 2003; Palter, et al., 2011; Pastor, Osman, Teitelbaum, Caty, & Langer, 2009; Williams & Carnahan, 2013). Cronbach’s α > 0.70 was selected as the predefined cutoff level for consistency. Items selected for inclusion into the final consensus list if at least 75 percent of the participants selected at least the "somewhat agree" (= 3) anchor on the final round of the survey.

Statistical analysis

Data collected for each round of the survey were analysed for internal consistency using Cronbach’s α. Medians (i.q.r) of panel members’ responses were calculated for all survey items. Responses to items on a categorical scale (yes/ no) were analysed as percentage responses. Data shown represents medians (i.q.r) or percentages unless stated otherwise. Statistical analysis was performed using a SPSS version 20.0 (IBM SPSS Statistics, IBM Corp., Armonk, NY, USA).

4.4 Results

Participants and survey rounds

A total of 19 major national and international organizations were identified as sources relevant to the identification of experts on the subject matter of interest. Furthermore, all members of the Advanced Training in Laparoscopic Abdominal Surgery (ATLAS) group were assessed for eligibility (Table 8). Thirty-three eligible experts agreed to participate in the Delphi survey and completed the first round; 25 of these (76 percent) continued on to complete the second survey round (Table 9). Geographic distribution of the panel was wide in both rounds (12 countries in round one and nine countries round two).
## Table 8: Organizations identified and screened for expert inclusion

<table>
<thead>
<tr>
<th>Organizations screened</th>
</tr>
</thead>
<tbody>
<tr>
<td>American College of Surgeons (ACS)</td>
</tr>
<tr>
<td>American Society for Metabolic and Bariatric Surgery (ASMBS)</td>
</tr>
<tr>
<td>Association for Surgical Education (ASE)</td>
</tr>
<tr>
<td>British Obesity and Metabolic Surgery Society (BOMSS)</td>
</tr>
<tr>
<td>Canadian Association of General Surgeons Association (CAGS)</td>
</tr>
<tr>
<td>Endoscopic &amp; Laparoscopic Surgeons of Asia (ELSA)</td>
</tr>
<tr>
<td>European Association for Endoscopic Surgery (EAES)</td>
</tr>
<tr>
<td>Federación Latinoamericana de Cirugía (FELAC)</td>
</tr>
<tr>
<td>German Society for Minimally Invasive Surgery (CAMIC)</td>
</tr>
<tr>
<td>German Society for Obesity and metabolic Surgery (CA-ADIP)</td>
</tr>
<tr>
<td>Indian Association of Gastrointestinal Endo Surgeons (IAGES)</td>
</tr>
<tr>
<td>International Federation for the Surgery of Obesity (IFSO)</td>
</tr>
<tr>
<td>International Federation of Societies of endoscopic surgeons (IFSES)</td>
</tr>
<tr>
<td>International Society of Laparoscopic Colorectal Surgery (ISLCRS)</td>
</tr>
<tr>
<td>Obesity Surgery Society of Australia and New Zealand (OSSANZ)</td>
</tr>
<tr>
<td>Italian Society for Endoscopic Surgery and New Technologies (SICE)</td>
</tr>
<tr>
<td>Society of American Gastrointestinal and Endoscopic Surgeons (SAGES)</td>
</tr>
<tr>
<td>Society of Laparoendoscopic Surgeons (SLS)</td>
</tr>
<tr>
<td>The Association of Laparoscopic Surgeons of Great Britain and Ireland (ALSGBI)</td>
</tr>
<tr>
<td>Geographical Region</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>North America</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Europe</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Australasia</td>
</tr>
<tr>
<td><strong>Total</strong></td>
</tr>
</tbody>
</table>

Table 9: Included expert panel

**Consensus and item selection**

Cronbach’s $\alpha$ for all survey items in the first round was 0.91, with four of the eight subsections calculated to have $\alpha$ values > 0.70 (Table 10). Free-text entries made by the panelists during the first round were used to generate 31 additional items for ranking in round two. In round two, Cronbach’s $\alpha$ was 0.92 for all items, and all subsections had $\alpha$ values > 0.70 (Table 10).
<table>
<thead>
<tr>
<th>Question ID</th>
<th>Question group</th>
<th>Cronbach (α)</th>
<th>Round 1 (n=33)</th>
<th>Round 2 (n=25)</th>
<th>Included</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>≥ 3 (%)</td>
<td>Missing data (n)</td>
<td>≥ 3 (%)</td>
<td>Missing data (n)</td>
</tr>
<tr>
<td>1.1</td>
<td>1 Definitions error</td>
<td>0.927</td>
<td>96.9 1</td>
<td>0.904</td>
<td>91.6 1</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td>51.5 0</td>
<td>78.3 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3</td>
<td></td>
<td>87.8 0</td>
<td>95.8 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td></td>
<td>42.4 0</td>
<td>45.9 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.5 a</td>
<td></td>
<td>88 0</td>
<td>4 [3-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6 a</td>
<td></td>
<td>92 0</td>
<td>4 [3.5-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.7 a</td>
<td></td>
<td>96 0</td>
<td>5 [4-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8 a</td>
<td></td>
<td>88 0</td>
<td>5 [3-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>2 Definitions technical error</td>
<td>N/A b</td>
<td>N/A 0</td>
<td>0.813</td>
<td>95.8 1</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>N/A 0</td>
<td>95.8 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3 a</td>
<td></td>
<td>95.8 1</td>
<td>4 [3.25-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.4 a</td>
<td></td>
<td>76 0</td>
<td>4 [2.5-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.5 a</td>
<td></td>
<td>76 0</td>
<td>3 [2.5-4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.6 a</td>
<td></td>
<td>91.6 1</td>
<td>4 [4-5]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7 a</td>
<td></td>
<td>84 0</td>
<td>4 [3-4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>3 Examples error unclassified</td>
<td>0.932</td>
<td>78.8 0</td>
<td>0.878</td>
<td>83.3 1</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>54.6 0</td>
<td>58.4 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td></td>
<td>84.8 0</td>
<td>79.1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td>84.8 0</td>
<td>83.3 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td>81.3 1</td>
<td>95.8 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.6</td>
<td></td>
<td>81.3 0</td>
<td>95.8 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td></td>
<td>72.7 0</td>
<td>78.3 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td></td>
<td>78.7 0</td>
<td>83.3 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question ID</td>
<td>Question group</td>
<td>Round 1 (n=33)</td>
<td>Round 2 (n=25)</td>
<td>Included</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cronbach (α) ≥ 3 (%)</td>
<td>Missing data (n)</td>
<td>Cronbach (α) ≥ 3 (%)</td>
<td>Missing data (n)</td>
</tr>
<tr>
<td>3.9</td>
<td></td>
<td>97 0</td>
<td>100 1</td>
<td>5 [5-5]</td>
<td>+</td>
</tr>
<tr>
<td>3.10</td>
<td></td>
<td>94 0</td>
<td>100 1</td>
<td>5 [4-5]</td>
<td>+</td>
</tr>
<tr>
<td>3.11a</td>
<td></td>
<td></td>
<td>75 1</td>
<td>3 [2.25-4]</td>
<td>+</td>
</tr>
<tr>
<td>3.12a</td>
<td></td>
<td>79.2 1</td>
<td>4 [3-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.13a</td>
<td></td>
<td>95.8 1</td>
<td>4 [3.25-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.14a</td>
<td></td>
<td>95.8 1</td>
<td>4 [3.25-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.15a</td>
<td></td>
<td>91.6 1</td>
<td>4 [3.25-4]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.16a</td>
<td></td>
<td>91.6 1</td>
<td>4 [3-4]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.17a</td>
<td></td>
<td>87.5 1</td>
<td>3 [3-4]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>3.18a</td>
<td></td>
<td>79.2 1</td>
<td>4 [3-4]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>4 Examples error suturing</td>
<td>0.921 84.8 0</td>
<td>0.834 87.5 1</td>
<td>4 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>4.2</td>
<td></td>
<td>42.4 0</td>
<td>45.8 1</td>
<td>2 [2-3]</td>
<td>-</td>
</tr>
<tr>
<td>4.3</td>
<td></td>
<td>36.4 0</td>
<td>34.8 2</td>
<td>2 [1-3]</td>
<td>-</td>
</tr>
<tr>
<td>4.4</td>
<td></td>
<td>77.8 0</td>
<td>75 1</td>
<td>4 [2.25-4.75]</td>
<td>+</td>
</tr>
<tr>
<td>4.5</td>
<td></td>
<td>57.6 0</td>
<td>75 1</td>
<td>3 [2.25-4]</td>
<td>+</td>
</tr>
<tr>
<td>4.6</td>
<td></td>
<td>66.7 0</td>
<td>83.3 1</td>
<td>3 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>4.7</td>
<td></td>
<td>78.8 0</td>
<td>83.3 1</td>
<td>3.5 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>4.8a</td>
<td></td>
<td>67 1</td>
<td>3 [2-4]</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4.9a</td>
<td></td>
<td>87.5 1</td>
<td>4 [3-4.75]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>4.10a</td>
<td></td>
<td>87.5 1</td>
<td>4 [3.25-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>Question ID</td>
<td>Question group</td>
<td>Cronbach (α)</td>
<td>≥ 3 (%)</td>
<td>Missing data (n)</td>
<td>Cronbach (α)</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
<td>--------------</td>
<td>---------</td>
<td>------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>5.1</td>
<td>5 Examples error clipping</td>
<td>0.628</td>
<td>N/A 0</td>
<td>0.913 1</td>
<td>91.7 1</td>
</tr>
<tr>
<td>5.2</td>
<td></td>
<td>N/A 0</td>
<td>100 1</td>
<td>4 [3-5]</td>
<td>+</td>
</tr>
<tr>
<td>5.3</td>
<td></td>
<td>N/A 0</td>
<td>70.9 1</td>
<td>3 [2-3.75]</td>
<td>-</td>
</tr>
<tr>
<td>5.4</td>
<td></td>
<td>N/A 0</td>
<td>87 2</td>
<td>4 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>6.1</td>
<td>6 Definitions event</td>
<td>N/A b</td>
<td>N/A 3</td>
<td>0.826 1</td>
<td>100 1</td>
</tr>
<tr>
<td>6.2</td>
<td></td>
<td>N/A 0</td>
<td>91.6 1</td>
<td>4 [3.25-4.75]</td>
<td>+</td>
</tr>
<tr>
<td>6.3a</td>
<td></td>
<td>N/A 0</td>
<td>95.7 1</td>
<td>4 [4-5]</td>
<td>+</td>
</tr>
<tr>
<td>7.1</td>
<td>7 Examples event</td>
<td>0.869 69.1</td>
<td>1 0.887 95.8 1</td>
<td>4 [3-4.75]</td>
<td>+</td>
</tr>
<tr>
<td>7.2</td>
<td></td>
<td>81.8 0</td>
<td>87.5 1</td>
<td>4 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>7.3</td>
<td></td>
<td>87.9 0</td>
<td>91.6 1</td>
<td>3.5 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>7.4</td>
<td></td>
<td>93.9 0</td>
<td>87.5 1</td>
<td>4 [3.25-5]</td>
<td>+</td>
</tr>
<tr>
<td>7.5</td>
<td></td>
<td>84.8 0</td>
<td>66.7 1</td>
<td>3 [2-4]</td>
<td>-</td>
</tr>
<tr>
<td>7.6</td>
<td></td>
<td>78.8 0</td>
<td>83.3 1</td>
<td>4 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>7.7</td>
<td></td>
<td>79.9 0</td>
<td>83.3 1</td>
<td>3 [3-4]</td>
<td>+</td>
</tr>
<tr>
<td>7.8</td>
<td></td>
<td>94 0</td>
<td>91.3 2</td>
<td>5 [4-5]</td>
<td>+</td>
</tr>
<tr>
<td>7.9a</td>
<td></td>
<td>91.7 1</td>
<td>4 [5-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7.10a</td>
<td></td>
<td>83.3 1</td>
<td>4 [4-5]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7.11a</td>
<td></td>
<td>75 1</td>
<td>3 [2.25-4]</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>7.12a</td>
<td></td>
<td>54.2 1</td>
<td>3 [2-3]</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7.13a</td>
<td></td>
<td>66.7 1</td>
<td>4.5 [3-5]</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>7.14a</td>
<td></td>
<td>50 1</td>
<td>2.5 [2-3]</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Question ID</td>
<td>Question group</td>
<td>Cronbach (α)</td>
<td>≥ 3 (%)</td>
<td>Missing data (n)</td>
<td>Cronbach (α)</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------</td>
<td>--------------</td>
<td>---------</td>
<td>-----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>7.15&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8 Weighting of scales</td>
<td>N/A&lt;sup&gt;b&lt;/sup&gt;</td>
<td>62.5</td>
<td>1</td>
<td>79.2</td>
</tr>
<tr>
<td>7.16&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
<td>87.5</td>
<td>1</td>
<td>5 [4-5]</td>
<td>100</td>
</tr>
<tr>
<td>7.17&lt;sup&gt;b&lt;/sup&gt;</td>
<td>N/A</td>
<td>79.2</td>
<td>1</td>
<td>4 [3-4]</td>
<td>100</td>
</tr>
<tr>
<td>7.18&lt;sup&gt;a&lt;/sup&gt;</td>
<td>N/A</td>
<td>39.1</td>
<td>2</td>
<td>4 [3-5]</td>
<td>100</td>
</tr>
</tbody>
</table>

Overall Survey 0.908 0.918

Table 10: Level of consensus for each survey round and final item selection

[N/A] no calculation since level of defined consensus not reached. [<sup>a</sup>] Additional items introduced in the 2nd survey round. [<sup>b</sup>] Scale has zero variance or too few data points for calculation of Cronbach's α.

Fifty-six (83 percent) of the items included in the second survey round that pertained to definitions and examples of error in laparoscopic surgery were ranked by at least “3” on the Likert-type scale by at least 75 percent of participants; thus, they were included in the reference list (Table 11). Of these, 51 items were rated above the cutoff threshold in both rounds, whereas four items (1.2, 3.7, 4.5, and 4.6) were initially below this cutoff in the first survey round but were rated more positively by participants in the second round. One item (7.5) that participants rated positively in round one subsequently received lower ratings by participants in round two and did not meet the threshold for inclusion (Table 10, Table 12). A total of 17 definitions and 39
examples of errors and events were ultimately selected for inclusion in the consensus list (Table 11). Twelve items were excluded based on the consensus criteria (Table 12).

The panelists felt that when using an error-rating system, observed errors should be weighted depending on error consequence (100 percent “yes”), assumed severity (96 percent “yes”), and whether they necessitate corrective action (80 percent “yes”).

<table>
<thead>
<tr>
<th>Item No.</th>
<th>Text description</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Act, assertion, or belief that unintentionally deviated from what is correct, right or true</td>
<td>American Heritage Dictionaries, 2011</td>
</tr>
<tr>
<td>1.2</td>
<td>Act or instance of deviating from an acceptable code of behaviour</td>
<td>American Heritage Dictionaries, 2011</td>
</tr>
<tr>
<td>1.3</td>
<td>A mistake or inaccuracy</td>
<td>American Heritage Dictionaries, 2011</td>
</tr>
<tr>
<td>1.5</td>
<td>Actions made due to lack of surgical knowledge, judgement, or technical skill</td>
<td>Expert</td>
</tr>
<tr>
<td>1.6</td>
<td>Action that deviates from acceptable safe practice, (intentionally or unintentionally)</td>
<td>Expert</td>
</tr>
<tr>
<td>1.7</td>
<td>A surgical act of commission or omission that could potentially result in an adverse event or outcome</td>
<td>Expert</td>
</tr>
<tr>
<td>1.8</td>
<td>Causing injury through lack of knowledge of laparoscopic physiology, technology, instrumentation and correct tissue planes</td>
<td>Expert</td>
</tr>
<tr>
<td>2.2</td>
<td>Errors where the operative technique was identified as the underlying cause of an adverse event</td>
<td>Regenbogen, et al., 2007</td>
</tr>
<tr>
<td>2.3</td>
<td>Using the wrong instrument or energy source for a task</td>
<td>Expert</td>
</tr>
<tr>
<td>2.4</td>
<td>Failure to identify anatomy precisely before acting</td>
<td>Expert</td>
</tr>
<tr>
<td>2.5</td>
<td>Pursuing an operation despite poor vision</td>
<td>Expert</td>
</tr>
<tr>
<td>2.6</td>
<td>Unsafe execution of an established technical exercise</td>
<td>Expert</td>
</tr>
<tr>
<td>2.7</td>
<td>Action that could potentially lead to tissue damage</td>
<td>Expert</td>
</tr>
</tbody>
</table>
### Definitions of the term *event*

**6.1** Unintended incidents in care that may result in adverse outcomes or may require additional care efforts to prevent adverse outcomes


**6.2** Consequences of errors requiring rectification or modification of initial plan

Expert

### Definition of the term *adverse event*

**6.3** Intraoperative event which is a result of a technical error and has a potential to lead to an adverse outcome if left uncorrected

Expert

### Examples of technical errors (unclassified)

| 3.1 | Overshooting when grasping and pulling on tissue beyond target | Video observation |
| 3.3 | Cutting tissue without lifting from underlying structures | Tang, Hanna, Joice, et al., 2004 |
| 3.4 | Pulling tissue in metallic trocar (swording) | Talebpour, et al., 2009 |
| 3.5 | Inserting trocar without view of tip | Pugh, et al., 2010 |
| 3.6 | Inserting suture passer/closure device without view of tip | Video observation |
| 3.7 | Blind insertion of instrument catching on omentum or falciform ligament | Video observation |
| 3.8 | Loss of pneumoperitoneum and retaining sharp instruments in abdomen | Video observation |
| 3.9 | Nonvisualization of tip of cautery (L-hook) whilst using | Video observation |
| 3.10 | Handling sharp/ hot instruments out of view or loss of orientation whilst insertion into abdomen | Video observation |
| 3.11 | Use of sharp dissection/ energy dissection in an area with poorly defined anatomy | Expert |
| 3.12 | Dissection in a wrong anatomic plane | Expert |
| 3.13 | Using the wrong instrument for a specific task | Expert |
| 3.14 | Cutting off target | FLS™ criteria/ Expert |
| 3.15 | Placing loop ligature off target | FLS™ criteria/ Expert |
| 3.16 | Inadequate force when manipulating tissue | Expert |
| 3.17 | Faulty trocar placement causing suboptimal view, ergonomy and/or surgical access | Expert |
| 3.18 | Stapling in an inadequate angel | Expert |

### Examples of errors in suturing

<p>| 4.1 | Bringing loaded needle out of view | Video observation |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4</td>
<td>Catching of nontarget tissue</td>
<td>Video observation</td>
</tr>
<tr>
<td>4.5</td>
<td>Throwing loop to tie a knot but not completing motion of an actual knot</td>
<td>Hamad, et al., 2007</td>
</tr>
<tr>
<td>4.6</td>
<td>Suturing through another loop of suture (unintentionally locking)</td>
<td>Video observation</td>
</tr>
<tr>
<td>4.7</td>
<td>Loss of grip on needle whilst driving through tissue, deviating from target</td>
<td>Video observation</td>
</tr>
<tr>
<td>4.9</td>
<td>Forceful extraction of the needle from the tissue</td>
<td>Expert</td>
</tr>
<tr>
<td>4.10</td>
<td>Criteria used in FLS (gap in knot, slippage of knot, gap between edges, knot disruption)</td>
<td>FLS™ criteria/ Expert</td>
</tr>
</tbody>
</table>

**Examples of errors in clip applying**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Clipping a structure without view of both tips of the clip</td>
<td>Video observation</td>
</tr>
<tr>
<td>5.2</td>
<td>Clipping in a pool of blood without view of underlying tissue</td>
<td>Video observation</td>
</tr>
<tr>
<td>5.4</td>
<td>Overlap of two clips</td>
<td>Ahlberg, et al., 2007</td>
</tr>
</tbody>
</table>

**Examples of events**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Bleeding after visualized error</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.2</td>
<td>Serosal tear</td>
<td>Hamad, et al., 2007</td>
</tr>
<tr>
<td>7.3</td>
<td>Hematoma of bowel/stomach wall or mesentery after handling</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.4</td>
<td>Liver laceration after trocar insertion or use of retraction</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.6</td>
<td>Bending/ breaking needle whilst suturing</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.7</td>
<td>Ripping suture whilst knot-tying</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.8</td>
<td>Diathermy/ energy source burn to nontarget tissue</td>
<td>Video observation</td>
</tr>
<tr>
<td>7.9</td>
<td>Unintended damage (mechanical or thermal) to the wall of a viscus/ organ</td>
<td>Expert</td>
</tr>
<tr>
<td>7.10</td>
<td>Failure of a stapler line</td>
<td>Expert</td>
</tr>
<tr>
<td>7.11</td>
<td>Any bleeding related to use of instruments including instruments with energy sources</td>
<td>Expert</td>
</tr>
<tr>
<td>7.16</td>
<td>Leaving a foreign body or piece of tissue</td>
<td>Expert</td>
</tr>
<tr>
<td>7.17</td>
<td>Performing an anastomosis under tension</td>
<td>Expert</td>
</tr>
</tbody>
</table>

**Table 11: Consensus list included items**

Origin of items: literature, expert opinion (comments in free-text fields), and video observations (research team). [FLS™] fundamentals of laparoscopic surgery.
<table>
<thead>
<tr>
<th>Item No.</th>
<th>Text description</th>
<th>Origin</th>
</tr>
</thead>
</table>

**Definitions of the term error**

1.4  If a planned sequence of mental or physical activities fails to achieve its intended outcome …..  J. T. Reason, 1990

**Examples of technical errors (unclassified)**

3.2  Slipping off tissue (e.g. whilst running the bowel or retracting gallbladder)  Video observation

**Examples of errors in suturing**

4.2  Dropping needle  Video observation

4.3  Picking needle up at the tip  Talebpour, et al., 2009

4.8  Breaking stitch  Hamad, et al., 2007

**Examples of errors in clip applying**

5.3  Dropping clips prematurely  Video observation

**Examples of events**


7.12  Any breakdown of imaging or gas insufflation  Expert

7.13  Twisting of bowel after anastomosis  Expert

7.14  Unintentional lock/ break of suture that requires a new start  Expert

7.15  Loss of needle/ clip/ gallstone requiring additional time for retrieval  Expert

7.18  Tissue ischemia following actions (clipping/ stapling/ dissection) that unintentionally cut off blood supply  Expert, Hamad, et al., 2007

Table 12: Consensus list excluded items

Origin of items: literature, expert opinion (comments in free-text fields), and video observations (research team)
4.5 Discussion

Surgical educators need to provide trainees with objective and meaningful feedback and critique on technical aspects of their work in order to improve performance. To date, there are no universal definitions of acceptable and poor technique in laparoscopic surgery. A reference list of such definitions was established in the present study through a Delphi consensus survey involving international experts. To facilitate usability, definitions of technical errors in laparoscopic surgery were complemented by examples of common errors and events.

*Surgeons as educators*

Technical skills learned throughout residency training are the result of years of observation and practice. This concept was challenged with the introduction of laparoscopic techniques since it was feared that due to a loss of tactile sensation, residents would no longer be able to adequately learn by assisting procedures as in open surgery (Morgenstern, 2005). Traditional training methods were therefore swiftly augmented through simulation techniques, with the benefit of simulation now being widely accepted and scientifically proven (Aggarwal, Moorthy, & Darzi, 2004; Zendejas, Brydges, Wang, & Cook, 2013). Nevertheless, much of surgical training still takes place in the live operating room. The initial concerns that trainees in laparoscopic surgery may not adequately learn through observation have recently been proven unsubstantiated (Nunobe et al., 2013). Therefore, surgical education should continue to focus on the role of the surgeon as a mentor and coach. This was pointed out by Kron (2011), who discussed the topic of surgical mentorship in his presidential address at the 91st Annual Meeting of the Association for Thoracic Surgery meeting, criticizing the fact that standardized teaching techniques are lacking, and emphasizing the difficulty in objectively defining surgical excellence. These concerns are valid in all surgical contexts but are of particular relevance in minimal access surgery. Minimal access surgery is undergoing continuous technical changes, with new approaches such as single incision and natural orifice surgery being developed and rapidly implemented into general practice. In this environment of modernization, it is important to promote basic surgical principles and safe surgical technique. The consensus list obtained in this
study facilitates these since it summarizes key principles regarding surgical technical errors that are considered of clinical relevance by international experts and opinion leaders in the field. The summary of classical examples can serve the surgeon educator as an easy-to-use reference guide for objective resident feedback or for use during task supervision.

*Delphi panel*

In the setting where little objective evidence is available to serve as a reference for the educator, consensus methods can be used to assist decision making. Combining the knowledge of a diverse group of individuals to create an applicable guideline is a common approach in consensus development methods (Murphy et al., 1998). However, panel composition is an important quality measure in Delphi surveys (Linstone & Turoff, 1975); therefore, selection criteria recommended by Delbecq et al. (1975) were followed. These included selecting individuals who represented decision makers, who would ultimately utilize and endorse the findings (Delbecq, Gustafson, & Van de Ven, 1975). Therefore, opinions of laparoscopic surgeons with an active role in surgical education were gathered in the present study. We believe these individuals represented opinion leaders and important stakeholders for future modifications in practice recommendations.

A further factor that may influence the quality of the consensus statements obtained is the size of the selected Delphi panel. To date, there are no evidence-based recommendations in this regard, and it has been suggested that the Delphi method should not rely on the panel to symbolize a representative sample for statistical purposes (Powell, 2002). Rather, the panel should consist of individuals – specialists trained and competent in the area of interest - who will ultimately utilize the results of the Delphi survey (Hsu & Sandford, 2007; Jones & Hunter, 1995). Panel sizes ranging from 10 to 50 participants reflecting a relatively heterogeneous group have been suggested to improve performance of the panel (Delbecq, et al., 1975; Murphy, et al., 1998; Witkin & Altschuld, 1995). In the present study, panel members were widely distributed geographically, reflecting different surgical cultures, practices, and subspecialities such as foregut, colorectal, and obesity surgeons as well as generalists without dedicated field of expertise beyond laparoscopic surgery.
Selection of items

The items selected for inclusion in the final consensus list were those for which at least 75 percent of participants at least "somewhat agreed" on the clinical and educational relevance of the item. This may appear to be a lenient criterion, but since the majority of the initial items were selected from examples of errors described in the literature and previously used in research and education, it can be assumed that these items had high content validity. We can also assume that all further items generated in the second round also have high content validity since they had been suggested by the panelists themselves (Goodman, 1987). Items not included, such as "slipping off tissue," were not deemed relevant by the panel despite having being linked to injuries in a prior analysis of error patterns in advanced laparoscopic surgery (Bonrath, Zevin, Dedy, & Grantcharov, 2013). The evidence supporting the relevance of this seemingly minor technical slip was obtained through objective observations of surgical performance, but the data were not available to the panel at the time of the Delphi survey. As a result, this example may well have appeared harmless to most surgeons and therefore not of clinical relevance.

Studies on structured objective analysis of error in laparoscopic surgery are necessary to deliver the much needed evidence-base for the design of practice recommendations. Several studies in this field have been successfully completed highlight the importance of errors in routine procedures (Joice, et al., 1998; Talebpour, et al., 2009; Tang, Hanna, Bax, et al., 2004; Tang, Hanna, Joice, et al., 2004). More work is needed to create a database of errors that have been shown to cause harm. Until such a database becomes available, expert judgment and opinion elicited in this study will represent the best level of evidence to support training recommendations.

Limitations

This study had several limitations. First, experts were recruited through affiliations with surgical societies. Although several large societies were screened, the list of analysed databases is far from inclusive. Nevertheless, analysis of the mentioned societies’ homepages revealed a high degree of redundancy regarding their members. This was especially true for European and Asian members, who in general were members of at least one large international society such as the
European Association for Endoscopic Surgery (EAES) and Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) as well as members of local societies. The majority of the expert panel were from the U.S., Canada, and Europe. Asia, South America and Africa were clearly underrepresented. This was not due to individuals from these regions not having been selected but rather more not wishing to participate in the study at the time. Irrespective of the geographic limitations, this study spanned three large regions representing a wide variety of different surgical training paradigms. Therefore, we believe that the results of this study are representative of current concepts in surgical education.

The effects of bias from not participating must also be discussed. Studies that explored the characteristics of nonrespondents contacted to participate in consensus generation revealed that respondents and nonrespondents were very similar (McKee, Priest, Ginzler, & Black, 1991). It was concluded that respondents who are on expert panels could be viewed as representative of their nonresponder colleagues. Similar results have been reported in studies that compared consensus generation in parallel groups, with a high level of agreement reported between the different expert groups (Duffield, 1993; Kastein, Jacobs, Van der Hell, Luttik, & Touw-Otten, 1993). These findings may be due to the criterion-based selection process of individuals rather than the methods of sampling as used in opinion surveys, since the individuals were all identified through similar characteristics. Delphi surveys, unlike classical surveys, do not rely on methodologies such as sampling or nonresponse analysis as the process of expert selection is in itself a biased technique that is common to all consensus generating techniques involving panels and advisory boards (Fink, Kosecoff, Chassin, & Brook, 1984; Murphy, et al., 1998).

The second limitation of this study was the attrition rate of 24 percent between the first and second rounds. This is common in Delphi surveys when multiple rounds of surveys are administered (Hsu & Sandford, 2007; Witkin & Altschuld, 1995). Attrition rates in two- and three-round Delphi surveys have been reported to be around 20 percent (de Villiers, de Villiers, & Kent, 2005; Mead, 1991). Using a modified Delphi approach, with a semistructured first round, we hoped to facilitate response and to encourage participation through all rounds. Initial analysis of responses revealed a high degree of overall agreement within the first round. We opted to conduct a second round of the survey in view of the multiple free-text entries that were made in the first round. Despite the attrition rate, answers provided by the participants in round two altered the result on the decision for or against inclusion in only five instances. Overall
consensus in the second round was as high as in the first round and therefore we feel the resulting consensus list to still reflect the overall opinion of the panel.

Finally, the Delphi process is not an exact science and does not lead to the detection of one “correct” answer. All results and statements need to be viewed with this in mind. It does not replace original research to generate evidenced-based data. On the other hand, the processes is very well suited for exploring opinions on topics on which there is little or no empiric data and which are subject to a high degree of controversy (Okoli & Pawlowski, 2004). The topic of this study was therefore well suited for exploration by Delphi technique. From an educator's standpoint, it is important that individuals entrusted with the training of the next generation of surgeons can relate to the opinions of peers and experts.

4.6 Conclusion

The list of technical errors in laparoscopic surgery established in this study can be used as a reference guide to increase standardization of training in laparoscopic techniques. Variations of standard techniques may be necessary in selected instances as demanded by the specific operative situation; therefore, adherence to this reference guide may not always be indicated. We advocate that this classification be used primarily as a guide for analyzing surgical performance with an error-focused approach, and as a reference for surgical educators to provide structured feedback to trainees.
CHARACTERIZING "NEAR MISS" EVENTS IN COMPLEX LAPAROSCOPIC SURGERY THROUGH VIDEO-ANALYSIS

This chapter describes a detailed analysis of error-event patterns as a means to explore injury mechanisms in routine LRYGB procedures. These examples deliver the evidence base for technical training principles in LRYGB, and serve as didactic material in technical training.

The contents of this chapter have been published in:

Characterising “Near Miss” Events in Complex Laparoscopic Surgery through Video-Analysis

Bonrath EM, Gordon LE, and Grantcharov TP

BMJ Quality and Safety, 6 May 2015

doi:10.1136/bmjqs-2014-003816

Copyright Article Author 2015, permission as per licensing agreement BMJ Journals
5.1 Abstract

*Background:* Root-cause analyses of surgical complications are of high importance to ensure surgical quality, but specific details on technical causes often remain unclear. Identifying subclinical intraoperative incidents attributable to technical errors is essential for developing rescue mechanisms to prevent adverse outcomes.

*Objective:* Descriptive study to characterize intraoperative technical error-event patterns in successful laparoscopic procedures.

*Methods:* Events (injuries) identified during prior blinded analyses of 54 unedited video recordings of bariatric laparoscopic procedures were subjected to a secondary review to determine the presumed underlying error mechanism. The recordings were obtained from one university-based bariatric collaborative program, and represented consultant, fellow, and shared trainee cases.

*Results:* Sixty-six events were identified in 38 recordings, whereas 16 videos showed no events. In 25 (66 percent) of the videos that showed events, additional rectification measures such as hemostasis or suture repair were required. Common identified events were minor bleeding (n=39, 59 percent), thermal injury to non-target tissue (n=7, 11 percent), serosal tears (n=6, 9 percent). Common error mechanisms were “inadequate use of force/distance (too much)” (n=20, 30 percent) and “inadequate visualization” during grasping/dissecting (n=6, 9 percent), “inadequate use of force /distance (too much)” using an energy-device (n=6, 9 percent), or during suturing (n=6, 9 percent). All events were recognized intraoperatively.

*Conclusion:* Analysis of successful operations allowed the identification of numerous error-event sequences. Reviewing injury mechanisms can enhance surgeons' understanding of relevant errors. This error awareness may aid surgeons in preparing for cases, help avoid errors and mitigate their consequences. Thus, this approach may impact future surgical education and quality initiatives aimed at reducing surgical risks.
5.2 Introduction

Numerous reports have explored the incidence of adverse outcomes and errors in healthcare. These reports provided valuable information on the root-causes of adverse outcomes and indicated that in surgery, the majority of errors occurred within the OR and that many of these errors can be considered technical in nature (A. A. Gawande, et al., 2003; S. O. Rogers, Jr., et al., 2006). Although the knowledge gained from these retrospective studies is valuable, information on errors that did not cause a visible complication is lacking. These “near misses”, (i.e. situations that had the potential to result in an injury or adverse outcome but failed to do so due to chance or through appropriate countermeasures, (National Patient Safety Foundation, http://www.npsf.org/for-healthcare-professionals/resource-center/definitions-and-hot-topics/patient-safety-dictionary-n-z), may not be captured by retrospective reviews of archived charts or malpractice claims, especially if the underlying cause is an intraoperative technical error, but are of high importance since they allow protective measures to be taken to avoid future adverse events before a clinically relevant injury manifests (Marella, 2007).

The nature of laparoscopy, which requires the transmission of an intraoperative picture to a screen, has enabled the thorough postoperative procedure analysis through video review. Sarker and colleagues (2005), for example, explored technical errors in laparoscopic cholecystectomy procedures performed by expert surgeons. Following their analysis, the authors devised a hypothetical model of “fluctuations of surgical errors” during operative cases. This model was based on the notion that minor errors occur throughout all cases. This “background technical error rate” was supported by the finding that even experts committed errors during routine cases, likely as a consequence of human nature and complexity of psychomotor skills required for laparoscopic surgery (Sarker, et al., 2005). Further studies have supported this hypothesis documenting varying error rates depending on procedure and level of training of the operating surgeon (Bonrath, Zevin, et al., 2013; Talebpour, et al., 2009; Tang, et al., 2005). As a result, the concept of an error focused training approach as a means to both reduce error and enhance error recovery through targeted training has been proposed (Darosa & Pugh, 2012; Meyerson, et al., 2012; D. A. Rogers, et al., 2002). This approach could benefit both trainees and practicing
surgeons since a recent analysis of surgical skill and patient outcomes in bariatric surgery revealed a significant correlation between surgeon skill and patient outcome, underlining the importance of optimized technical competence at all levels of surgical training (Birkmeyer, et al., 2013).

In the case of trainees, it is acknowledged that they will invariably commit numerous mistakes, and that this could be considered a valuable source of knowledge for experiential learning, provided there is a system facilitating identification, analysis and constructive feedback (Dror, 2011; Wu, et al., 1991). Therefore, the first step towards error reduction and mitigation appears to be to foster an understanding of common errors and to create awareness of potential injury mechanisms by acknowledging error-event patterns (Meyerson, et al., 2012). The objective of the present study, therefore, was to characterize common intraoperative error-event mechanisms in routine complex laparoscopic procedures.

## 5.3 Methods

*Study design*

This study was conducted as a secondary review of data, involving the retrospective review of procedural video clips.

*Error analysis*

All error analyses were conducted using the GERT (Bonrath, Zevin, et al., 2013). The GERT represents a simplified step-independent framework that can be used to categorize technical errors in laparoscopic surgery (Bonrath, Zevin, et al., 2013).
Definitions

In the present study, “errors” represent the smallest unit of deviation from the intended operative course. Based on the GERT framework, errors could be described by the type of deviation “inadequate use of force or distance (too much or too little)”, “wrong orientation of instrument or dissection plane”, and “inadequate visualization” as well as by the task during which they were observed (e.g. during abdominal access, use of retraction, use of energy devices, grasping and dissection, cutting, transection and stapling, clipping, suturing, suctioning) (Bonrath, Zevin, et al., 2013). The events in the present study correspond to “minor events” as detailed in prior classifications potentially requiring actions but without affecting the overall course of the operation (Champion, Meglan, & Shair, 2008). The term “error mechanism” in the present study refers to the error that was identified as contributing to the analyzed event.

Study sample

Ethics boards have approved several research projects involving error analysis of procedural videos, which have been conducted in a second to second analysis by a bariatric surgeon, in a blinded fashion. The video analysis software (Studiocode V.5, Sportstec Ltd., Warriewood, Australia) used allows for time-stamped marking of defined instances (in our case errors and events) on a video-independent timeline. Within the analysis software, the GERT analysis was combined with the procedural step classification of the hierarchical task analysis underlying the BOSATS tool to enable an accurate association of an error within a procedural step (Zevin, et al., 2013). Video sequences of the time-stamped events had been extracted with the datasets. These sequences included the marked event together with the preceding time frame that included any marked errors. The original videos had been deleted on completion of the primary analyses, retaining only the anonymized video segments which were the basis for the present secondary analysis.

For the present study, all data sets of LRYGB procedures (n=54) reflecting 78 hours of operating time, were assessed to identify those that had been marked with intraoperative events. The primary video pool of the data set represented unedited routine cases from one university based multisite bariatric collaborative program, and was a mix of "single bariatric surgeon only" (n=19)
and "shared trainee" (n=35) cases. Shared cases involved more than one surgeon acting as the primary surgeon for parts of the procedure, with any combination of a specialist bariatric surgeon, advanced minimally invasive surgery fellow, or trainee participating.

Determining event mechanisms

The aforementioned video segments were re-reviewed by the original rater, and reviewed by a second rater, a fellowship-trained practicing bariatric surgeon and educator (blinded to surgeon identity and training level). Both raters independently assessed the video segments describing the presumed event and stating which error they felt had caused the observed event. Following the methodology detailed by Regenbogen et al. (2007) descriptions of events and underlying error-mechanisms were compared between both raters. In cases of disagreement the two raters discussed until consensus was achieved. Inter-rater agreement was calculated based on the results of the initial independent review.

Statistical analysis

This was a descriptive study with the variable of interest being error-event sequences. Inter-rater reliability was assessed using Cohen’s Kappa. All analyses were performed using SPSS V.20 statistical software (IBM®).
5.4 Results

Sixty-six events were identified in 38 of the 54 videos. Sixteen videos showed no events. In 25 (66 percent) procedures that contained at least one event, a rectification intervention (Table 13) was required. The median rectification duration was 111 seconds (range 6-820 seconds).

| **Bleeding** | To stop bleeding, surgeons could apply clips to the tissue to occlude the bleeding vessel or apply electrocautery to coagulate it |
| **Serosa tears** (tear in the outer layer of the bowel wall) | To reinforce the damaged bowel wall, surgeons can oversew the injury |
| **Devascularized bowel edges** (poor blood supply to resection boarders) | In the setting of visibly reduced blood flow (ischemia) viability of the tissue may be compromised. If clearly compromised, the segment needs to be removed. This resection usually involves a few centimeters of bowel, and is carried up to the point where the bowel wall appears healthy and well perfused. |

**Table 13: Examples of rectification measures**

**Inter-rater agreement**

Inter-rater reliability for event description was excellent (Cohen’s Kappa =0.977, p<0.001, 95% CI 0.932-1) and inter-rater reliability for underlying error mechanism was substantial (Cohen’s Kappa =0.79, p<0.001, 95% CI 0.685-0.897) (Landis & Koch, 1977).

**Consensus**

Consensus discussion was required for one event: rater one assessed the injury as a hematoma of the bowel wall, whereas rater two described the injury as a serosa tear. After review of the clip and discussion, both raters agreed that the best injury description of the injury was a serosa tear (with secondary bleeding). With regards to error mechanisms, the raters disagreed in 12
instances. In four of these, the disagreement was with regards to error task-group (e.g., grasping versus use of energy depending on which instrument was presumed to have caused the injury); in eight, the presumed underlying error mode differed (e.g., too much versus too little force).

Events

The 66 observed events were instances of hematoma and minor bleeding (n=39), thermal injury to non-target tissue (n=7), serosa tear (n=6), devascularization of a small bowel segment after jejunal transection (n=5), entangled or broken sutures (n=4), torn falciform ligament (n=2), perforated mesentery (n=1), staple line failure (n=1), and non-target tissue caught in staple line (n=1).

Error mechanisms

The most commonly observed error mechanisms in the present sample were due to “use of inadequate force or distance (too much)” when grasping/dissecting (n=20) which could be seen to lead to avulsion or tearing of tissue which could lead to bleeding or serosal injuries, “inadequate visualization” during grasping/dissecting (n=6) commonly associated with injury to vessels in the not visualized deeper tissue layers, “inadequate use of force or distance (too much)” when suturing (n=6) resulting in either overshooting and inadvertent minor vessel injury beyond the target, or ripping of sutures requiring a new-start, “inadequate use of force or distance (too much)” or “wrong orientation of instrument or dissection plane” whilst using an energy device (n=6, and n=5, respectively) often leading to thermal injury of non-target tissue, while “inadequate visualization” whilst using an energy device (n=4) mostly resulted in bleeding. “Wrong orientation of the instrument or dissection plane” during stapling/transecting (n=5) was commonly seen to lead to devascularization of a segment of the transected small bowel. Less frequent were errors due to “wrong orientation of instrument” during suturing (n=4), or during clipping (n=1), “inadequate use of force or distance (too much)” while stapling/transecting (n=3) or clipping (n=1), “inadequate use of force or distance (too little)” while grasping (n=2) or stapling/ transecting (n=1), “inadequate visualization” during stapling/ transecting (n=1), and suturing (n=1).
Steps | Description
--- | ---
Splitting omentum | The omentum may be divided to facilitate tensionless positioning of the bowel during the creation of the gastrojejunal anastomosis
Measuring biliopancreatic limb | Starting at the ligament of Treitz the small bowel is measured distally to a length of approximately 40-60 cm and divided
Creating the jejunostomy | Following the division of the bowel, the alimentary limb is measured out (75-150 cm). The biliopancreatic limb is reconnected to the alimentary limb at the measured out point. The reconnection of the two bowel segments is termed the jejunoojejunostomy
Gastric pouch creation | The stomach is partially divided in a horizontal angle at the level of the lesser curve. From this division line, the stomach is then transected in a vertical fashion up to the angle of His to result in a small residual stomach pouch
Gastrojejunostomy creation | The new stomach pouch is reconnected to the bowel (alimentary limb/Roux-limb). This new connection is termed gastrojejunostomy
Mesenteric closure | The defect between the two transected mesentery edges is sutured closed to prevent internal herniation of bowel

**Table 14: Overview of Operative steps of LRYGB**

The steps described can be performed in a varying order and represent only a basic overview of the procedure. The steps of measuring biliopancreatic limb, jejunostomy, gastrojejunostomy or pouch creation can be further subdivided into several smaller steps as detailed in a prior hierarchical task analysis (Zevin, et al., 2013).

**Enactors and procedure step**

The majority of events in the present sample were caused by practicing bariatric surgeons (n=41), whereas a fewer number were caused by fellows (n=15) and trainees (n=10). An overview of the main operative steps of the LRYGB procedure is shown in Table 14. The events were most frequently observed during the operative step of gastric pouch creation (n=30), with the majority (n=18) of these occurring during the steps related to dissecting the lesser curve and creating a posterior gastric tunnel in preparation for transverse gastric division. The steps relating to creating the GJ (n=11), measuring the biliopancreatic limb (n=11), and creating the JJ (n=5) accounted for the majority of remaining events. Other steps combined, such as adhesiolysis,
splitting omentum, positioning the Roux-limb, mesenteric closure, only accounted for nine events; however some of these steps were not performed in all procedures. No events were observed during the introduction of trocars and set-up phase, or during the closure phase of the procedure.

5.5 Discussion

Although surgeons strive to consistently adhere to sound surgical principles and wish to avoid technical errors, the present study demonstrated that numerous error-event sequences may still be identified in successful routine operations since human error can never be fully avoided. The most common injuries were due to basic surgical tasks and were predominantly enacted by trained surgeons. The present analysis can thus inform surgeons and trainees alike about potential hazards and remind surgeons that even the most basic task execution may have an adverse consequence.

Numerous studies have been conducted to identify factors that may contribute to adverse patient outcome (Birkmeyer, et al., 2013; Cottam et al., 2007; Krell et al., 2014; Leape et al., 1991; Reames et al., 2014; Regenbogen, et al., 2007; S. O. Rogers, Jr., et al., 2006). These studies focus on patient outcome as a function of an identified and diagnosed complication. Instances where a potential hazard was recognized and mitigated remain outside the scope of these reports leading to a significant loss of data having clinical, educational and economical relevance. Subsequently, a few prospective studies have been conducted with the aim to better understand these “near miss” situations in surgery, either through direct observation, (de Leval, et al., 2000) or by applying video recording techniques (Guerlain et al., 2005; Hu, Arriaga, et al., 2012). Hu et al. (2012) and de Leval et al. (2000) applied a comprehensive approach to analyzing perioperative events, and demonstrated that events, even those that could potentially endanger patient safety, are very common. Furthermore, a cumulative effect of even minor events was noted to adversely impact patient outcome (de Leval, et al., 2000). These studies, however, did not primarily assess the impact of factors associated with technical skills. The present work,
focused on technical aspects of surgical performance. As technical errors have been identified as a major source of injury resulting in disability (Cottam, et al., 2007; Regenbogen, et al., 2007), knowledge about intraoperative risks and appropriate rescue mechanisms may have a significant impact on patient safety and clinical outcomes. For example, common technical maneuvers such as blunt, blind dissection of the lesser curve of the stomach during pouch creation were identified as a leading cause of bleeding. Although all episodes of bleeding were minor and managed promptly, additional efforts were frequently required to achieve adequate control. Where one surgeon may easily control bleeding, another may make inappropriate decisions leading to a cascade of adverse events that could result in an adverse outcome. Therefore, by acknowledging the hazard in the step, surgeons can mentally prepare and assign cognitive resources for potential error recovery mechanisms (Dror, 2011). Similarly, by better understanding the origin of events, specific maneuvers can be avoided altogether. For example, missed enterotomies have been identified as a significant source of adverse patient outcome leading to lawsuits in bariatric surgery (Cottam, et al., 2007). By acknowledging the fact that a serosal tear, as a precursor to an enterotomy, may be the result of inadequate application of force during grasping the bowel, surgeons should be more vigilant and wary of any handling of bowel off-screen as potential injuries may be missed. Although most bariatric surgeons are likely to be aware of the risk of handling bowel with too much force, the risk of too little force may be underestimated as a recent expert consensus revealed (Bonrath, Dedy, Zevin, & Grantcharov, 2013b). It must be noted, that not all events identified in the present sample should be classified as “near misses” as they are unlikely to endanger patient safety. Incidents such as breaking or entangling sutures, torn falciform ligament and catching of non-target tissue in staple lines would likely only disrupt the flow of the procedure and require extra efforts to rectify. Furthermore, in the current sample the majority of required rectifications were only minimal as was reflected by the short duration of these interventions. These events could be classified as “minor events” (Champion, et al., 2008), rather than “near misses”. Similar results to the present study were found in previous work, although the distribution of error mechanisms differed (Bonrath, Zevin, et al., 2013). In addition, the video sample for the present study also included recordings from several different primary surgeons and hospital sites, thus generalizing the findings to include discreet variations in surgical techniques.
Furthermore, although the analysis was focused solely on technical errors, the role of knowledge and judgment deserves consideration. Frequently, a technical error may be the starting point of an error-event-complication cascade, but whether an event is recognized, interpreted correctly and, subsequently managed appropriately lies in the domain of knowledge and judgment. In addition, several errors, such as applying a stapler in the wrong orientation across the bowel, although technical in execution are commonly the consequence of misinterpretation of anatomy or lack of procedure knowledge.

All surgeons are prone to committing errors, and numerous studies have highlighted that errors are frequent (Bonrath, Zevin, et al., 2013; Sarker, et al., 2005; Talebpour, et al., 2009; Tang, et al., 2005; Tang, Hanna, Joice, et al., 2004); subsequently all surgeons have the potential to cause an inadvertent injury. In the present study, the majority of events were caused by practicing surgeons, which is likely due to the sample of videos used for analysis. Nineteen of the original videos were single surgeon cases; in addition in shared procedures, trainees would infrequently have performed two of the three steps with the most events (pouch creation and GJ creation). These two steps are more frequently performed by consultant level surgeons or fellows. This is in part due to the prevailing belief that the steps are more complex and that events may be harder to rectify. The current data do support the notion, that these steps may possibly be considered more difficult. Further objective analysis of error distributions using error-event ratios in combination with procedural step analysis is currently underway to document procedure step complexity in LRYGB.

The OR represents an essential environment for learning which cannot be fully replaced. In the context of experiential learning, the opportunity to learn from errors represents a valuable source of information that can be used to teach surgical decision making, risk management and error recovery mechanisms. But it may not be necessary for the individual to learn from mistakes made by themselves, it may also be sufficient to learn from exemplar errors. Subsequently, recent training concepts have evolved around topics of error training, aiming at instructing error recognition, error rescue and risk management (Darosa & Pugh, 2012; Dror, 2011; Meyerson, et al., 2012; D. A. Rogers, et al., 2002). Although the specific error mechanisms and events identified in the present study may not be generalizable to other surgical settings using variations in surgical technique, educators can develop a knowledge base of common technical errors relevant to their specific setting by conducting video reviews of routine cases. This knowledge
could subsequently be used to design targeted technical training interventions to address the most common consequential errors. Edited video segments can be used as educational material highlighting frequent injury mechanisms, which may aid trainees in understanding why specific maneuvers are undesirable. By reducing errors that result in events, procedure duration and costs may be reduced since event rectification may require additional materials and time efforts depending on the nature of the injury. Error preparedness may thus help in reducing overall costs; a topic that will need to be explored in future research.

**Limitations**

The current study has several limitations. First, the video clips used represented a convenience sample and, therefore, the event rate of 70 percent cannot be generalized. Potentially, in samples that were not selected by the primary surgeon, event rates could well be higher. However, the interest was not in determining the event frequencies. It was rather to characterize error-event mechanisms and to demonstrate feasibility of the method. Due to the heterogeneity of the original study sample total error counts were not a focus of this study, and all errors that had not resulted in an event (inconsequential) were not part of the present study. Second, since participation in the research studies was voluntary, contributing surgeons were free to select which videos to submit from their personal educational archives. This may have resulted in surgeons selecting only cases deemed to show good performance, which may be the reason why more significant injuries were not observed. Rare but remarkable and dangerous events have been described in the context of case reports (Higa, Szomstein, & Rosenthal, 2012; Huerta, Li, & Livingston, 2006). However, less critical events have largely not been the topic of publications and these seemingly “routine” events that can commonly occur should also be known by surgeons and trainees. Third, no information on patient outcome was available; consequently, the relevance of the identified events for patient outcome could not be ascertained. Nevertheless, due to the nature of some of these injuries and the potential for harm if left unrectified, several of these minor events still qualify as “near misses” warranting investigation. A prospective observational study is currently underway to determine the relevance of intraoperative errors and events on patient outcome, as well as to assess the relationship between technical errors and other measures such as nontechnical performance and environmental factors. Last, due to the
anonymous retrospective nature of the data, specific feedback was not afforded to the individual surgeons. The value of routine video recording in healthcare has recently been emphasized (Makary, 2013); the current study highlights how even seemingly uncomplicated cases can be a useful source of education, and that timely feedback may be valuable to surgeons of all levels of training.

5.6 Conclusion

Error awareness is essential in daily surgical practice and surgical training. The current study highlights the benefits of detailed video analysis to create a database of common injury mechanisms and video clip repository that can be used in tailoring future training interventions. In addition, practicing surgeons should be encouraged to review their own operations as even successful procedures can be a valuable source of learning prompting self-reflection. Understanding the casual relationship between minor errors and intraoperative events is essential for the development of effective error rescue mechanisms for future cases.
This chapter details the randomized controlled trial conducted to determine the effectiveness of surgical coaching in bariatric surgery. The evidence gathered support the relevance of coaching as a means to structure experiential learning in the OR. The approach combines observational assessment methodologies with evidence-based teaching content which have been discussed in the previous chapters.

The contents of this chapter have been published in:

*Comprehensive Surgical Coaching Enhances Surgical Skill in the Operating Room: A Randomized Controlled Trial*

Bonrath EM, Dedy NJ, Gordon LE, and Grantcharov TP
Annals of Surgery 2015 [Epub ahead of print]


The use of the manuscript is by permission of the copyright holder.
6.1 Abstract

**Objectives:** The aim of the study was to determine whether individualized coaching improved surgical technical skill in the operating room to a higher degree than current residency training.

**Background:** Clinical training in the operating room is a valuable opportunity for surgeons to acquire skill and knowledge; however, it often remains underutilized. Coaching has been successfully used in various industries to enhance performance, but its role in surgery has been insufficiently investigated.

**Methods:** This randomized controlled trial was conducted at one surgical training program. Trainees undergoing a minimally invasive surgery rotation were randomized to either conventional training (CT) or Comprehensive Surgical Coaching (CSC). CT included ward and operating room duties, and regular departmental teaching sessions. CSC comprised performance analysis, debriefing, feedback, and behaviour modeling. Primary outcome measures were technical performance as measured on global and procedure-specific rating scales, and surgical safety parameters, measured by error count. Operative performance was assessed by blinded video analysis of the first and last cases recorded by the participants during their rotation.

**Results:** Twenty residents were randomized, 18 completed the study. At post-training the CSC group (n=9) scored significantly higher on a procedure-specific skill scale compared to the CT group (n=9) [median 3.90 (i.q.r. 3.68-4.30) versus 3.60, (2.98-3.70), \( P = 0.017 \)], and made fewer technical errors [10 (7-13) versus 18 (13-21), \( P = 0.003 \)]. Significant within-group improvements for all skill metrics were only noted in the CSC group.

**Conclusion:** Comprehensive Surgical Coaching enhances surgical training and results in skill acquisition superior to conventional training.
6.2 Introduction

Surgery is a craft (Traynor, 2011), which is mastered through extensive training with the goal of acquiring sufficient knowledge and skills to maintain a consistently high level of performance. Although in recent years there have been significant investments in technology and innovative curricula, training during clinical encounters has not changed and valuable learning opportunities remain underutilized. The mechanisms of learning in the OR can be described with the experiential learning theory as popularized by Kolb (1984). Experiential learning requires the active involvement of the trainee in an experience with subsequent reflection and critical analysis. The process aims to identify new strategies derived from the reflection which can be applied to successfully master future similar experiences. Despite wide recognition of the experiential learning theory, the educational approach has been criticized as being "unguided" or "unstructured" (Kirschner, et al., 2006).

Evidence suggests that training in the OR is less purposeful and occurs infrequently (Scallon, Fairholm, Cochrane, & Taylor, 1992), and although trainees are provided with ample opportunities to engage in practical experiences, learning may be limited by a lack of analytic reflection on these experiences, due to inaccurate self-assessment (D. A. Davis, et al., 2006). Methods to improve the self-assessment processes, such as debriefing, have been suggested, but their implementation remains a challenge (M. Ahmed, et al., 2013).

Experiential learning in the OR could be enhanced by a structured approach including objective assessment, structured debriefing, feedback, behaviour modeling, and guided self-reflection. This is commonly employed in competitive sports. In coaching, a coach can follow a structured methodology to guide the coachee to achieve self-determined goals (Grant, 2011; Palmer, 2011). The aim of the present study was to assess the effectiveness of an instructional approach, "Comprehensive Surgical Coaching", using structured feedback, debriefing, and behavioural modeling as a means to enhance experiential learning in the OR and improve surgical skill and clinical safety.
6.3 Methods

**Trial design**

This was a randomized Controlled Trial (Clinicaltrials.gov registration: NCT01875679) with two treatment arms, comprehensive surgical coaching (CSC) or conventional residency training (CT). The allocation ratio to each treatment was 1:1.

**Participants**

Surgical residents and fellows from a single large General Surgery training program (University of Toronto) undertaking a MIS rotation with a minimum duration of two months during the study period (2013-2014) were recruited. All trainees, who routinely acted as primary surgeon (under supervision) for steps of the JJ creation of LRYGB procedures, and had recording equipment in the OR were eligible for enrollment. Institutional ethics approval was obtained and informed consent was obtained from all participants before study enrollment.

**Index procedure**

For the purpose of this study, only the steps required to complete the JJ were analyzed. The steps from “elevating the colon to identify Treitz” up to “closure of the enterotomy defect” were attributed to the “creation of the jejunojejunostomy”. Participants completed a checklist questionnaire after each procedure detailing the steps they performed as a primary surgeon, as well as any instances of supervisor takeover or case sharing.

**Interventions**

Participation in this study did not influence any aspect of clinical residency duties. Although staff surgeons supervising participants during their MIS rotations were informed of the study at
the beginning of the study period, they were blinded to resident participation status, as well as to
group allocation in case of participation. To achieve this, staff surgeons were not actively
involved in the study (recruitment or training), and participants were asked to not disclose their
participation or their group allocation to staff or colleagues. Furthermore, study activities
occurred removed from the surgery, outside the clinical setting, not influencing daily residency
routines, and thus could be kept discrete and confidential.

Conventional training

Participants in the CT group underwent regular teaching sessions offered by the Department of
Surgery and had responsibilities on the ward and in the OR. All participants had exposure to
advanced laparoscopic procedures, under the supervision of a staff surgeon during their rotation.
Frequency of teaching and feedback in the OR during the cases was determined by the
responsible staff surgeon during the rotation.

Comprehensive Surgical Coaching

In addition to conventional residency training, coaching sessions were structured around a
modified PRACTICE coaching model (Palmer, 2011). Instructional methods used to address
deficiencies and enhance positive behaviours during the coaching sessions were feedback,
debriefing with video-debriefing, and behaviour modeling using both good and poor examples.
To determine individual training needs, video review of the operative steps performed by the
participants was conducted. In the coaching sessions, participants received individualized
formative feedback on the basis of the structured assessments, and developed generalizable
training goals together with the coach. Debriefing was conducted using video playback of edited
clips, encouraging self-reflection of observed performance. After identifying performance
deficiencies, practical solutions to address those and enhance future performance were
developed, communicated and implemented by the participant during subsequent cases. The
effectiveness of the modified performance was discussed in the following coaching session. A
flexible approach was used depending on the participant’s individual needs, and instructions
were scaffolded with less guidance towards the end of the two-month study participation. As coaching represents a personalized approach, distinct learning goals and objectives were identified for each participant. Therefore, initial skill level only played a minimal role and coaching could be focused around prior existing knowledge and adjusted throughout the sessions as required to accommodate differences in learning pace.

Coaching was delivered by a surgeon-educator with fellowship training in bariatric surgery and extensive experience in technical skills analysis. The training interventions were planned to be administered on a weekly basis, although variability in coaching frequency was allowed to accommodate resident schedules and operating exposure. The coaching was performed outside the clinical operative environment.

**Outcome measures**

***Baseline and post-training assessment***

From all submitted recordings, two videos per participant were used for formal assessment: the first procedure performed and recorded was used for the baseline (BL), and the last procedure during the enrollment period was used for the post-training (PT) assessment.

**Primary outcome measures: Technical performance**

Surgical skill was as assessed using the OSATS global rating scale (Martin, et al., 1997). Procedure-specific skill was measured using BOSATS tool (Zevin, et al., 2013). Error rate was measured using the GERT (Bonrath, Zevin, et al., 2013).

**Secondary outcome measures:**

**Perceived usefulness:** Participants of the CSC group completed a PT questionnaire to assess their subjective perception of the usefulness of the coaching intervention.
**Self-reflection:** Each participant completed one self-assessment of operative performance at the end of the study enrollment using the BOSATS and OSATS scales. Participants’ self-assessed scores were measured against scores from the blinded assessors to determine whether CSC improved the accuracy of the surgeon's self-assessment of technical performance.

**Time-commitment:** The feasibility of the intervention was evaluated by recording the time for preparation and delivery.

**Sample size**

An a priori power calculation on the basis of results of prior studies applying the BOSATS (Zevin, 2014; Zevin, et al., 2013) and GERT (Bonrath, Zevin, et al., 2013) frameworks was performed. The aim of the present study was to increase operative performance (global, procedure-specific skill) and reduce technical errors. The study was powered to detect a skill enhancement, on all parameters, from an “untrained” level to a “trained” level (within group), and to detect a difference between training modalities (between groups). For this, scores for untrained and trained and expected scores for conventionally trained residents were established using data from the aforementioned studies. Depending on metric and comparison mode between five and eight participants were required in each group to obtain a power of 0.8 with an α value of 0.05 for a two-tailed test. The targeted recruitment to account for “drop-outs” was set at 24, with an interim analysis conducted once 16 participants had been recruited. Data analysis was conducted in blocks. Recruitment was continued until completion of the interim analysis.

**Randomization**

Web-based blocked randomization with a block size of six and an allocation ratio of 1:1 was used. Blocked randomization was used to ensure equal participant distribution over the continuous recruitment period (Altman & Bland, 1999; Kang, Ragan, & Park, 2008). The
randomization plan was used to create 24 sequential, sealed allocation envelopes which were used by the researcher during enrollment. Recruitment occurred in the order the participants became eligible for enrollment. The researchers involved in recruitment and enrollment were blinded to the allocation algorithm using only the sealed envelopes.

**Data analysis**

Video recordings showing only the unlabeled intra-abdominal view of the procedure (laparoscopic camera view) were de-identified with regard to participant identity, and list randomized in blocks to allow for the interim analysis. The BL and PT videos of each analysis block (interim and final) were mixed with a large number of randomly selected recordings from all sections of the participants’ learning curves to ensure rater blinding and to minimize contrast effects between BL and PT performances. Overall, a total number of 114 recordings were reviewed. Global ratings (OSATS and BOSATS) were assessed in duplicate by two raters to minimize rater subjectivity, while GERT assessments were performed by only one rater as this tool has shown to deliver highly reliable scores with regards to total error counts requiring little interpretation (Bonrath, Zevin, et al., 2013). For the GERT ratings intra-rater reliability was determined for 14 recordings from the entire pool of submitted videos which were randomly selected and reviewed twice with a 6 to 8 week interval between ratings.

OSATS scores were reported as total scores, BOSATS as mean scores including only steps performed by the participant, thus ranging from 1 (lowest) to 5 (highest). The number of technical errors for steps performed at BL and PT were assessed as total sum, and subsequently corrected for operative substeps performed both at BL and PT for each individual. The substeps of the procedure were defined according to the hierarchical task analysis underlying the BOSATS tool (Zevin, et al., 2013).

**Statistical analysis**

Descriptive statistical analysis was performed. Normality of distribution was assessed by the Shapiro-Wilk test. For data that was not normally distributed, group comparisons were
conducted by non-parametric statistics, using the Mann-Whitney $U$ test for between-group comparisons, and the Wilcoxon-rank sign tests for within-group. Spearman correlations were used to examine self-versus blinded assessments, and the relationship between case number and primary outcome measures. Normally distributed data (change scores) were assessed by the $t$-test for independent samples. Inter-rater agreement was assessed by Intra-class Correlation using a two-way mixed model for absolute agreement. Intra-rater agreement was calculated using Intra-class correlation using a one-way random model. ICCs are reported for average and single measures, with 95 percent confidence intervals (CI). All other data are reported as median (i.q.r) unless stated otherwise. Categorical data were compared by the Fisher exact test. Data analyses were performed using the IBM® SPSS® 20.0 package (IBM Corp., Armonk, NY).

6.4 Results

Participants

Twenty participants were recruited and randomized; one was discontinued during the study (CSC group) and one was excluded from analysis (CT group) (Figure 7, CONSORT diagram (Moher, Schulz, & Altman, 2001)). Participant demographic data are shown in Table 15.
<table>
<thead>
<tr>
<th></th>
<th>CT (n=9)</th>
<th>CSC (n=9)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>31 (28-35)</td>
<td>34 (30-34.5)</td>
<td>0.346</td>
</tr>
<tr>
<td>Postgraduate year</td>
<td>4 (3-4.5)</td>
<td>4 (3.5-5)</td>
<td>0.328</td>
</tr>
<tr>
<td>Sex, male : female</td>
<td>5:4</td>
<td>8:1</td>
<td>0.294*</td>
</tr>
<tr>
<td>Handedness, right : left</td>
<td>9:0</td>
<td>9:0</td>
<td>1.0*</td>
</tr>
<tr>
<td>Prior MIS rotations, months</td>
<td>4 (1-5.5)</td>
<td>3 (2-7)</td>
<td>0.877</td>
</tr>
<tr>
<td>Prior exposure to simulation, yes : no</td>
<td>8:1</td>
<td>8:1</td>
<td>1.0*</td>
</tr>
<tr>
<td>Advanced MIS procedures assisted, n</td>
<td>3</td>
<td>0</td>
<td>0.088*</td>
</tr>
<tr>
<td>&lt;10 cases</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>10-50 cases</td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>&gt;50 cases</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Advanced MIS procedures performed as surgeon, n</td>
<td>4</td>
<td>3</td>
<td>1.0*</td>
</tr>
<tr>
<td>&lt;10 cases</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>10-50 cases</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>&gt;50 cases</td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Table 15: Demographic details**

Advanced MIS cases: for example bariatric cases (excluding gastric bands), colorectal, pancreatic, or gastric surgery; [MIS] Minimally Invasive Surgery; [*] Nominal data reported as counts, P-values from the Fisher exact test for nominal data. All other data (continuous and ordinal) reported as medians (interquartile range), P-values from the Mann-Whitney U test.

Correlations between prior operative exposure, training level, and months in minimal invasive surgery rotations with skill parameters (OSATS, BOSATS, and error count) were not significant at BL (Table 16).
<table>
<thead>
<tr>
<th></th>
<th>OSATS</th>
<th>BOSATS</th>
<th>Error count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Postgraduate year</td>
<td>$r_s = -0.016, P = 0.949$</td>
<td>$r_s = -0.069, P = 0.787$</td>
<td>$r_s = 0.148, P = 0.558$</td>
</tr>
<tr>
<td>Completed MIS rotations, months</td>
<td>$r_s = -0.059, P = 0.817$</td>
<td>$r_s = 0.054, P = 0.830$</td>
<td>$r_s = 0.184, P = 0.466$</td>
</tr>
<tr>
<td>Prior exposure as surgeon</td>
<td>$r_s = -0.022, P = 0.931$</td>
<td>$r_s = 0.033, P = 0.897$</td>
<td>$r_s = 0.198, P = 0.430$</td>
</tr>
</tbody>
</table>

**Table 16: Correlation between baseline demographic data and baseline skill parameters**

Correlation using Spearman Rho ($r_s$) between baseline demographic data, such as training level (postgraduate year), prior rotations and prior surgical experience, and baseline performance metrics for the entire group (n=18). [OSATS] Objective Structured Assessment of Technical Skill; [BOSATS] Bariatric Objective Structured Assessment of Technical Skill.

*Concurrent surgical exposure and number of coaching sessions*

During the MIS rotations, all participants were exposed to a variety of surgical procedures, either as the operating surgeon or the assistant. There was no significant difference between groups in the distribution of cases performed and assisted, or in the number of JJ’s performed throughout the study period (Table 17). Participants in the CSC group completed a median of 4 (3.5-5) coaching sessions.
Figure 7: CONSORT 2010 Flow Diagram (Moher, et al., 2001)
### Table 17: Concurrent surgical exposure during a two month rotation

Comparison of procedures performed and assisted during a two month rotation. Classification of procedures [advanced]: for example, bariatric cases (excluding gastric bands), colorectal, pancreatic, gastric surgery; [intermediate]: fundoplications, hiatal hernias; [basic]: diagnostic laparoscopy, cholecystectomy, appendectomy, inguinal hernia. Nominal data reported as counts (number of participants), *P*-values from the Fisher exact test, [#] Jejunojejunostomies reported as medians (interquartile range), with *P*-value from the Mann-Whitney U test. [CT] Conventional Training; [CSC] Comprehensive Surgical Coaching; [MIS] Minimally Invasive Surgery.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Cases</th>
<th>CT (n)</th>
<th>CSC (n)</th>
<th><em>P</em>-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced MIS procedures performed as surgeon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>0</td>
<td>2</td>
<td>0.471</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Advanced MIS procedures assisted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>5</td>
<td>3</td>
<td>0.637</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Intermediate MIS procedures performed as surgeon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>8</td>
<td>9</td>
<td>1.0</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Intermediate MIS procedures assisted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>9</td>
<td>9</td>
<td>1.0</td>
</tr>
<tr>
<td>Basic MIS procedures performed as surgeon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>2</td>
<td>0</td>
<td>0.576</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Basic MIS procedures assisted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>8</td>
<td>6</td>
<td>0.576</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Open cases performed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10</td>
<td></td>
<td>5</td>
<td>5</td>
<td>1.0</td>
</tr>
<tr>
<td>10-50</td>
<td></td>
<td>4</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>&gt;50</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Jejunojejunostomy performed during study period</td>
<td>9</td>
<td>(7-10.5)</td>
<td>10 (7.5-14.5)</td>
<td>0.416*</td>
</tr>
</tbody>
</table>
Primary outcome measures

Technical performance

Inter-rater agreement for the 36 BL and PT videos was excellent for OSATS (ICC average measures, 0.950; CI, 0.902-0.975); ICC single measures, 0.905; CI, 0.822-0.951) and BOSATS (ICC average measures, 0.940; CI, 0.882-0.969; ICC single measures, 0.887; CI, 0.789-0.940). Intra-rater reliability for error count was excellent (ICC single measures, 0.90; CI, 0.729-0.967). All subsequent analyses for OSATS and BOSATS were performed using the average scores from both raters.

Between-group comparison at BL showed no significant differences for the three main performance metrics (OSATS score, BOSATS score, and error count) (Table 18). At the end of the study, PT results differed significantly between groups, with higher technical skill scores achieved in the CSC group as measured on the BOSATS tool, and lower technical error counts. OSATS scores had improved in both groups, but differences between groups did not reach significant levels (Table 18). Within-group comparison for the CT group showed no significant improvement of any of the three skill parameters, whereas the CSC group had improved significantly on all metrics (Table 18).

In addition, change scores (PT-BL) were calculated for OSATS scores, BOSATS scores and technical error counts. Change scores were normally distributed and analyzed by independent sample t test. Change scores were significantly higher in the CSC group for both ΔOSATS [mean, 7.11; SD, 4.78 versus mean, 2.22; SD, 4.27; t(16)=2.29, \( P = 0.036 \) d=1.08, and ΔBOSATS [mean, 0.68; SD, 0.35 versus mean, 0.10; SD, 0.44; t(16)=3.07, \( P =0.007 \) d=1.45. Δtechnical error was significantly lower in the CSC group [mean, -6.33; SD, 4.98 versus mean, 3.11; SD, 6.13; t(16)= -3.59, \( P =0.002 \) d=1.69.]
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CT</td>
<td>CSC</td>
<td>P (Between Groups)</td>
</tr>
<tr>
<td><strong>OSATS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>20 (17.5-25.5)</td>
<td>21.5 (17-22.75)</td>
<td>0.531</td>
</tr>
<tr>
<td>PT</td>
<td>24 (19.5-26.5)</td>
<td>27 (25.75-30.75)</td>
<td>0.065</td>
</tr>
<tr>
<td><em>P value (within group)</em></td>
<td>0.137</td>
<td><strong>0.008</strong>*</td>
<td></td>
</tr>
<tr>
<td><strong>BOSATS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>3.45 (3.05-3.63)</td>
<td>3.3 (3.06-3.60)</td>
<td>0.948</td>
</tr>
<tr>
<td>PT</td>
<td>3.60 (2.98-3.70)</td>
<td>3.90 (3.68-4.30)</td>
<td><strong>0.017</strong>*</td>
</tr>
<tr>
<td><em>P value (within group)</em></td>
<td>0.438</td>
<td><strong>0.008</strong>*</td>
<td></td>
</tr>
<tr>
<td><strong>Technical Error</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BL</td>
<td>15 (7-21)</td>
<td>16 (11.5-20.5)</td>
<td>0.649</td>
</tr>
<tr>
<td>PT</td>
<td>18 (13-21)</td>
<td>10 (7-13)</td>
<td><strong>0.003</strong>*</td>
</tr>
<tr>
<td><em>P value (within group)</em></td>
<td>0.313</td>
<td><strong>0.008</strong>*</td>
<td></td>
</tr>
</tbody>
</table>

**Table 18: OSATS, BOSATS scores, and error counts between and within groups**

Results shown as medians and interquartile ranges. OSATS and BOSATS values represent averaged scores of both blinded raters. BOSATS scores represent scale means with 1 being the lowest and 5 the highest value achievable. OSATS scores represent scale totals with 7 being the lowest and 35 the highest score achievable. *P* values for within group comparisons: Wilcoxon signed ranks test. *P* values for between group comparison: Mann-Whitney U test. Significant values flagged with [*]. [BL] Baseline; [PT] Post-Training; [CT] Conventional Training; [CSC] Comprehensive Surgical Coaching group.

**Influence of case volume**

To evaluate the relationship between case number and performance, correlations between each of the three outcome measures and case number were conducted. All correlations in the CSC group were significant (OSATS: \( r_s = 0.62, P = 0.006 \); BOSATS: \( r_s = 0.65, P = 0.004 \); error count: \( r_s = - \).
0.55, \( P = 0.018 \)). None of the correlations were significant in the CT group (OSATS: \( r_s = 0.31, P = 0.217 \); BOSATS: \( r_s = 0.25, P = 0.315 \); error count: \( r_s = 0.16, P = 0.537 \)).

Staff takeover

At BL, staff takeover was required in six instances (CT, n=3; CSC, n=3; \( P = 1.0 \)), in four of these cases staff take-over resulted in the participant not being afforded an opportunity to attempt one or more substeps of the procedure. The durations of take-over did not differ between CT and CSC groups [128 (range, 109-464) s vs. 430 (178-1057) s, \( P = 0.40 \)]. At PT, staff take-over was required in two instances only in the CT group; in one instance the participant subsequently did not attempt the substeps "positioning limbs on stapler, ensuring symmetric alignment" and "firing stapler".

Secondary outcomes

Perceived usefulness: All participants in the CSC group found the coaching sessions very useful. Participants felt that the video demonstrations were very useful (n=8) or useful (n=1). Moreover, all individuals thought coaching very strongly (n=7) or strongly (n=2) influenced their performance in the OR. All agreed very strongly that structured formal as well as informal feedback should be integrated in routine training.

Self-reflection: Correlation of self-assessment scores on BOSATS and OSATS scales with scores obtained by blinded video review did not correlate in the CT group for mean BOSATS (\( r_s = -0.448; P = 0.265 \); n=8) or OSATS (\( r_s = 0.46; P = 0.247 \); n=8), whereas in the CSC group, self-assessed scores were significantly correlated with expert assessment for both mean BOSATS (\( r_s = 0.78; P = 0.013 \); n=9) and OSATS (\( r_s = 0.85; P = 0.004 \); n=9).
Time-commitment: For 33 of the 40 coaching sessions, records of time investment required to prepare and deliver the coaching intervention were available. The median time required for viewing and structured assessment of the procedures was 40 (30-50) minutes, video-editing was 16 (10-23) minutes, and defining learning objectives was 7 (4-9) minutes. The median time to deliver the coaching intervention was 25 (23-28) minutes.

6.5 Discussion

The “Comprehensive Surgical Coaching” approach detailed in this study involved approximately four coaching sessions per participant in a two month timeframe and resulted in significant improvement of surgical skill and safety metrics, with large effect sizes for all primary outcome measures. Furthermore, the skill improvement exceeded the improvement achieved through conventional residency training in two of the metrics. In addition a correlation between skill improvement and cases performed was only noted in the CSC group. The intervention was very well received with all participants advocating for implementation in the residency program. The time commitment was minimal, with a coaching session requiring approximately one and a half hours to be prepared and administered. Notably, after two months of structured coaching, participants in the CSC group, but not in the CT group, were able to more accurately assess their own performance, a prerequisite for future continuing professional development.

Index procedure

Considerations during the selection of the index procedure for the present study included several requirements: the procedure had to be common, standardized, and have a high level of complexity in order to warrant dedicated training and facilitate the observation of errors and lower skill levels. Bariatric surgery fulfills all these criteria. LRYGB has become a common
procedure (M. M. Davis, Slish, Chao, & Cabana, 2006), leading to a high resident exposure. Furthermore, the procedure is deemed technically challenging, with the learning curve estimated to encompass 75 to 100 procedures (Oliak et al., 2003; Schauer, Ikramuddin, Hamad, & Gourash, 2003), and with lower surgical skill shown to correlate with adverse patient outcome (Birkmeyer, et al., 2013). Recent studies confirmed that supervised active resident participation in LRYGB did not lead to higher complication rates (Iordens, et al., 2012; Rovito, et al., 2005), supporting the integration of the procedure into residency training. Only the first half of the LRYGB procedure, fashioning the jejunojejunostomy, was selected for this study, as trainees frequently perform these steps, which are common to a variety of laparoscopic procedures involving the creation of an entero-enteroanastomosis.

Coaching in surgery

The present trial is the first clinical study to document the impact of comprehensive coaching in a real OR environment using performance outcome parameters.

High-impact publications have suggested adoption of concepts from other high performance industries such as coaching in professional sports as alternative approaches to continuing medical education and surgical training (A. Gawande, 2011; Greenberg, et al., 2014). “Coaching” has been used in numerous contexts including sports, business, relationship, and general life coaching (Ives, 2008). Although coaching has many facets, and may be grounded in multiple psychological constructs, several characteristics are common to a broad range of different approaches. Examples include the administration of coaching in a systematic process, individualization to the end user, use of listening and questioning, and goal-orientedness (Ives, 2008). The coaching framework selected for this study, PRACTICE was used to guide the overall structure of coaching sessions (Palmer, 2007, 2011). The advantage of coaching in this setting was that the personalized approached allowed for the training of participants exhibiting a variety of differing skill levels at BL. Irrespective of training level, coaching could be used to maximize the learning for each individual through formulating level appropriate realistic goals. A similar approach has recently been reported using the GROW framework in a simulated laparoscopic environment (Singh, Aggarwal, Tahir, et al., 2014). A video-based coaching approach was trialed at the level of practicing staff, and was found to be a feasible and valuable
learning opportunity, although performance outcome measures were not reported (Hu, Peyre, Arriaga, Osteen, et al., 2012).

The instructional strategies employed during the coaching sessions in the present study involved debriefing, feedback, and behaviour modeling using video modeling. Debriefing, described as “facilitated or guided reflection in the cycle of experiential learning” (Fanning & Gaba, 2007)[pg 116], represents a vital component to enhance the learning experience (Lederman, 1984). Following Kolb's framework (Kolb, 1984), debriefing can be instrumentalized to enhance learning at all experiential learning stages (Walker, 1984). For the present study debriefing was implemented in a standardized fashion, with all debriefings facilitated by one researcher. In contrast to classic debriefing as an activity that immediately follows an experience such as a simulation (Lederman, 1984), the sessions in the present study were conducted removed from the actual OR which allowed for a structured analysis of performance before to the session and enabled the preparation of demonstrative material. Furthermore, by allowing time to pass between the actual experience and the coaching session, participants were afforded an opportunity to self-reflect and allow potential anxiety or stress sentiments to subside prior to revisiting the performance.

Feedback represents another essential instructional concept in medical education as it enables the clarification of any discrepancies between actual and desired performance (Ende, 1983). The relevance of feedback to enhance laparoscopic skill acquisition has been demonstrated in several studies in a simulated environment (Paschold, et al., 2014; Strandbygaard, et al., 2013), and has been shown to improve surgical skill in the OR after live observation (Grantcharov, et al., 2007) and when combined with video-debriefing (Hamad, et al., 2007). In the present study feedback followed the initial debriefing phase of the coaching sessions and was used to emphasize good performance and point out deficiencies. Even though debriefing and feedback represent separate instructional methodologies, they are frequently combined in a fluid process to enhance learning (M. Ahmed et al., 2012).

Behaviour modeling builds on the concept that individuals learn through behavioural observations reflecting on outcomes of those observed behaviours (Bandura, 1977). In the present study, video-modeling using augmented feedback (N. J. Hodges, et al., 2003) was combined with video vignettes of exemplar behaviour. Video-feedback describes the method of
playing back the individual's own performance and has been shown to enhance learning of laparoscopic skills (Hamad, et al., 2007; Nakada, et al., 2004). In the present study, time requirement during coaching sessions was minimized by editing the participant's performance to highlight only the most relevant learning points. In addition, exemplar video clips obtained from video-analyses of trainee and staff cases were used to demonstrate tasks required for successful completion of the procedure as well as to demonstrate potential injury mechanisms. The use of both good and poor examples was aimed at promoting generalizability of modeling to the complex environment (Baldwin, 1992; Taylor, Russ-Eft, & Chan, 2005).

Skill metrics in context

Because skill levels have been identified to correlate with patient outcome (Birkmeyer, et al., 2013) it is important to view the results in the broader context of surgical performance. However, comparability between studies may be limited because of variations in rating methodologies used. Thus, it is important to view the results within contextual data obtained in comparable fashion within a similar cohort. Data from a cohort of graduating resident from the same training program, with equivalent exposure to MIS rotations during their training, and assessed by the same rater were available for comparison (Zevin, 2014). OSATS scores from the graduating cohort were slightly lower although comparable (mean 23.6, SD 5.7) to the results recorded for the conventionally trained cohort at the end of their rotation in the present study. Although the CT group also improved in skill, none of the within-group improvements were statistically significant. Thus, results from both cohorts suggest that the levels of skills that can be achieved through conventional residency training in this setting appear to range in the good mid-field of the OSATS scale. In contrast the CSC group achieved higher OSATS scores, with a significant within-group improvement indicating a substantial benefit through the additional training intervention. Interestingly, throughout the study an error reduction was only observed in the CSC group, while within the CT group errors increased. This suggests that approaches in current training do not focus on error reduction. Reducing technical errors, is of high relevance for patient safety as technical errors have been shown to be a major source of injury and adverse outcome (Regenbogen, et al., 2007), and should therefore be targeted in modern surgical training concepts (Darosa & Pugh, 2012).
Influence of case volume

At BL there were no differences in demographics between groups, and remarkably, no correlations between BL skill parameters (OSATS, BOSATS and error count) and prior operative experience, months on MIS rotations or post graduate training level were found. This finding underscores that experiential discovery learning with little structure or guidance, as is common in the operative environment (Roberts, et al., 2009), may result in inconsistent learning results. Furthermore, as the study was embedded in regular residency training, as would be expected participants had varying amounts of operative encounters. Although none of these differences achieved statistical significance, it may be argued that the results observed were due to the CSC group performing more JJ cases. Although case number was correlated significantly with outcome metrics in the CSC group, with more exposure resulting in improved performance, this effect was not observed in the CT group. This finding reiterates that that without the additional structured feedback and reflection on clinical performance; clinical encounters per se may not automatically result in improved performance.

Feasibility of coaching sessions

This trial was designed with a pragmatic attitude aiming to show effectiveness in the setting of current residency training (Thorpe et al., 2009). The interventions were planned to be administered on a weekly basis, but this was frequently not possible and coaching sessions were most commonly performed on a biweekly basis. The sessions required approximately one and a half hours to prepare and administer, while the time commitment for the resident was only around 25 minutes. The coaching sessions were kept deliberately brief in order to not disrupt the resident's busy day. All participants in the CSC group strongly agreed that this type of intervention should be included in residency training. When considering a coaching program, several factors need to be acknowledged. First, the time requirements of coaching sessions depend on the length of the procedure, the number of trainees and frequency of sessions. Second, whether coaching is to be implemented as a continuous feature of training or rather will be performed only within specific rotations or intervals. Last, who will be conducting the coaching, staff surgeons or dedicated educators? These factors will influence how coaching is implemented into training and clinical practice. In the present study, the focus was on only one group of
trainees, those performing LRYGB. The time investment observed for each participant approximated 1\(\frac{1}{2}\) hours for each session. Most coaching was conducted on a biweekly basis. In addition, during the study it was noted that as trainees progressed and became familiar with coaching, sessions became less directive and trainer input became less frequent. This finding would suggest that periods of coaching could be mixed with periods of self-directed learning, but this structure will need to be tested in future implementation studies. Furthermore, in the present study coaching was conducted based on an advanced procedure (LRYGB). The guiding principle was the choice of a procedure that would allow repetitive exposure as a primary surgeon. Thus, coaching could also be arranged for more junior trainees structured around common procedures such as laparoscopic cholecystectomies.

Relevance of improved Self-assessment

The present study demonstrated that the coaching intervention involving structured and guided self-reflection, helped participants to become more accurate self-assessors. Self-assessment is a skill which is a prerequisite for successful continuing medical education and ongoing professional development because the valid and objective assessment of a physician’s knowledge and respective limitations is an important factor in self-regulation in the health-care professions (Gordon, 1991). Unfortunately, evidence to date suggests that physician's self-assessment abilities are lacking, and thus are frequently inaccurate when referenced against external measures (D. A. Davis, et al., 2006). After two months of coaching, the participants in the CSC group were able to conduct self-assessments of their performance more accurately than the CT group even at the lower end of the skill scale as measured against ratings from a blinded rater.

Limitations

The present study had several limitations because of the pragmatic nature of the trial. First, the training targeted skills improvement in an educational environment; therefore, an impact on patient outcome was not assessed. Focusing on patient outcome in educational studies as primary or secondary outcome measures has recently been criticized (Cook & West, 2013). Because
residents are in a supervised learning environment, their actions may not be directly correlated with patient outcomes since multiple factors may also play a role (Cook & West, 2013). For example, in the present study, residents were continuously supervised in the OR, and any poor performance would have led to the staff taking-over the case and rectifying the situation. Second, instructional methods that have individually been shown to benefit learning in various contexts were combined; therefore, it cannot be differentiated how much each methodology contributed to the overall effect. Nevertheless, the instructional approaches detailed are frequently combined despite representing distinct entities. Last, it must be discussed that the results were drawn from one large training program, with interventions delivered by one coach, thus limiting the generalizability. Further studies involving multiple sites and multiple coaches will need to verify whether the methodology translates in other settings.

6.6 Conclusion

Experiential learning in the OR can be augmented through a structured coaching approach, employing concepts of debriefing, feedback and behaviour modeling. This methodology resulted not only in improved technical skill levels and reduction of errors but also enhanced the participants’ self-assessment skills. Comprehensive Surgical Coaching represents an effective and efficient intervention to enhance current residency training and potential applications in continuing professional development.
7
GENERAL DISCUSSION

This chapter summarizes the findings of the previously described studies and discusses the most relevant points pertinent to surgical training in the OR in a competency-based learning environment.

The primary aim of this PhD-project was to design a training intervention for surgical residents and fellows that structured the experiential learning processes within the OR to improve the trainee’s surgical technical skill. The training intervention was organized around coaching frameworks and was based on performance analysis, individualized feedback, debriefing and behaviour modeling as strategies to augment the learning processes surrounding experiences made in the OR. The results of the randomized controlled trial reported in Chapter 6 support the rejection of null hypotheses formulated in Chapter 2 of this thesis. The alternative hypotheses that the coaching intervention would improve the participant’s technical skills as measured by global rating scales and reduced error counts were supported by the results.

Before these results could be achieved, several separate steps needed to be completed. In laparoscopic surgery, assessment methodologies for identifying technical errors and frameworks of definitions guiding acceptable surgical task execution were lacking. Further evidence-based examples for teaching were absent. For this reason, several research objectives were identified that had to be addressed before the design of the comprehensive surgical coaching program was possible. The studies describing how each research objective was accomplished are detailed in chapters within this thesis, each delivering a portion of the scientific basis for the coaching intervention. The results of these studies are discussed in detail within each of the chapters, but I will also summarize the global findings in this general discussion. Furthermore, I will discuss the most relevant points identified throughout this research project that need to be considered when analyzing the surgical learning environment and when planning to introduce structured training interventions to guide the experiential learning processes in the OR in order to avoid discovery learning.
Chapter 3 described the design of the GERT, which was based on the results of a prior literature review examining error assessment methodologies (Bonrath, Dedy, et al., 2013a). This study helped identify shortcomings in existing frameworks that informed the design of the new assessment tool. The new instrument was structured around an existing framework that had been developed outside the field of health-care (Embrey, 1986), and that had been adapted for use in surgery (Joice, et al., 1998). This established framework called "observational clinical human reliability analysis" (OCHRA), although well developed and supported by several sources of validity, relies on the surgeon’s adherence to a particular procedural step order. OCHRA assesses errors based on appropriate step completion (step order, steps missing, steps repeated) and based on task execution (executional errors) (Joice, et al., 1998; Tang, Hanna, Joice, et al., 2004). This step-dependent analysis is reliant on a thorough procedure analysis. The procedure analysis is performed by a panel of experts and involves the decomposition of the operation into component tasks and steps. Once a set of steps and tasks have been defined, the panel members determine the ideal order in which these tasks are to be completed (A. Cuschieri & Tang, 2010). As OCHRA can only be applied once the process of procedure deconstruction has been accomplished, the method is not readily generalizable. Furthermore, OCHRA requires adherence to a specific “ideal” step order, substantially limiting variability in procedure execution. This step adherence is problematic in advanced complex procedures, where numerous minor technical variations have been established. In addition, expert surgeons may not adhere to a predefined order of steps during an operation, as they may have developed their own specific approach, which may affect scoring using task and step specific scales (Beard, et al., 2007). Therefore, in the design of the GERT, the concept of error modes derived from the human reliability analysis approach of OCHRA was modified to include only executional deviations (considered technical errors), that could lead to an error of task execution. To obtain a better description of the errors, error modes were also linked to nine generic surgical tasks deemed applicable to laparoscopic surgery: abdominal access; use of retractors; use of energy devices; grasping and dissection; cutting, transection and stapling; clipping; suturing; use of suction; and other unclassified (Bonrath, Zevin, et al., 2013). The GERT framework thus describes technical errors with regards
to how a motion was considered erroneous, and with regards to the task during which the erroneous moves were observed. This detail was deemed important, as to enhance the formative value of the error analysis by embedding the error mode in a meaningful context. For the trainee, feedback detailing tasks that contained the most errors, combined with descriptions of erroneous motions, represents meaningful information. This information can be used to focus attention during subsequent cases, and can serve as a basis for deliberate practice protocols to address the most commonly observed errors.

The validity evidence gathered during the development of the GERT was reported according to the older framework. The results though, can freely be transcribed into the new unified framework. According to the unified framework, validity evidence sources from content, response process, internal structure, and relationship to other variables were obtained (AERA, et al., 1999). Consequential evidence was not gathered during the development of the GERT, but as performance was modified as a consequence of structured analyses that included the use of the GERT in the coaching intervention, this would suggest that the tool may have consequential effects due to its value in formative feedback.

Chapter 4 described the results of an expert consensus initiative aimed at obtaining clinically relevant definitions of erroneous performance in laparoscopic surgery. The chapter also detailed the concerns regarding inconsistencies in terminology and understanding of appropriate task execution. In order to facilitate objective assessment, assessors need to apply the same terms and rules. Similarly, as much of training occurs within the OR environment, content is guided by the knowledge base of the surgeon in the role of the educator. As shown in Chapter 1 of this thesis, teaching behaviours greatly vary, and teaching is rarely standardized. For that reason, it was of importance to obtain a reference guide of terms and examples to aid educators and assessors in judging performance and in establishing an evidence-base for intraoperative teaching. In a field where empiric evidence is largely lacking, consensus methodologies represent useful alternatives that may aid in generating reference guides. The Delphi method is a well-established consensus methodology that has frequently been used in health-care for this purpose (Murphy, et al., 1998). Using the Delphi approach consensus was achieved on a significant number of topics pertaining to definitions and examples of errors and events encountered in laparoscopic surgery. The list of included items can serve as a basis for the assessment of skill using the GERT, and can help educators determine which maneuvers are considered erroneous and should be addressed in
teaching. This standardization may aid in ensuring that trainees in different educational environments can relate to a set of common principles deemed relevant by international leaders in laparoscopic surgery. As all items that had initially been included in the Delphi survey were definitions and examples that had been used in prior studies, or were derived from direct observations or expert comments, the excluded items still remain clinically relevant, yet not all educators shared this opinion. Therefore, the consensus guide will need to be supplemented with empiric evidence as this becomes available. For this purpose a thorough review of injury mechanisms was conducted as detailed in Chapter 5, with the aim to verify specific potentially dangerous maneuvers. Results of both the consensus study and the analysis of injury mechanisms were used to guide teaching during the coaching intervention detailed in Chapter 6.

Chapter 5 details the review of identified intraoperative events. The events identified in prior video analyses represented “minor events” as detailed in earlier classifications (Champion, et al., 2008). These were defined as events, potentially requiring remedial actions, but without affecting the overall course of the operation. Since the events did have the potential to cause harm though, they could also be termed “near miss” situations (i.e. situations that had could result in an injury or adverse outcome but failed to do so due to chance or through appropriate countermeasures (National Patient Safety Foundation, http://www.npsf.org/for-healthcare-professionals/resource-center/definitions-and-hot-topics/patient-safety-dictionary-n-z). Identifying these “near miss” situations is of high importance, since understanding their causation will allow for the introduction of protective measures to avoid future adverse events before a clinically relevant injury manifests (Marella, 2007). This principle is of high importance in surgical training, as residents must be made aware of potential hazards, so that they can learn to avoid them or learn how to deal with them if they do arise. Thus, the information gained in Chapter 5 informed much of the didactic content used in modeling during the coaching intervention described in Chapter 6.

Chapter 6 addressed the primary aim of this PhD project. The evidence gathered during the prior studies was used to design a training intervention relevant to teaching residents and fellows the advanced laparoscopic procedure of LRYGB using their performance as a basis for the structured training. Instructional methods, shown to be effective in the OR (feedback, debriefing, behaviour modeling), were combined and structured around a personalized coaching framework. The training approach was successful in improving the trainees' skill level and improving the accuracy of self-assessments. The results of this trial are discussed in-depth in the discussion.
section of Chapter 6, and will for this reason not be repeated here. From the summarized general findings, several topics for discussion emerged, that will be discussed in more detail in the following subsections.
Comprehensive, frequent, objective assessment is central to modern competency-based training. The most important feature of in-training assessments is to allow for meaningful, constructive feedback. Therefore, an assessment must be sufficiently detailed. Global rating scales have been shown to deliver valid and reliable results, and reflect general skill competencies required to complete procedures successfully. Since they are intended to be generalizable though, they frequently lack detail due to vague wording used in the anchors (K. Ahmed, et al., 2011). Procedure-specific scales allow for a more detailed analysis through the assessment of a number of individual steps. These procedure-specific scales either use global ratings or binary ratings to express the degree to which the steps were completed appropriately. Both systems allow for the identification of steps that need further attention during training and allow for meaningful formative feedback (Beard, et al., 2007; Larsen, et al., 2008). These two scale systems (global and task-specific scales) assess one dimension of technical skill, the way in which a task is deliberately executed to achieve a specific goal. Error analysis is complementary to this type of skill assessment, as it enables a detailed review of the tasks that deviated from optimum performance. Maneuvers such as over- or undershooting, using instruments in the wrong plane, or without adequate visualization are not appropriately captured by global or task-specific observational tools, which aim to describe an “overall” impression of performance. Error analysis conversely looks at each move in detail; the sum of these moves usually forms the basis of global assessments. Thus, for formative purposes ideally all three modalities should be combined to obtain a meaningful, comprehensive representation of performance. During the coaching intervention reported in Chapter 6, three scales were combined to obtain a comprehensive assessment of performance on all dimensions of technical skill. The global assessments were made using the well-established OSATS scale; task-specific analysis was conducted using the BOSATS scale; and error counts were obtained using the GERT. All three measurements have sufficient validity evidence to support the interpretation of the results in the context in which the measurements were made. This objective comprehensive review was then used to structure the subsequent coaching session identifying the most pertinent general learning goals for each individual session. The results of the assessment were not shared with the trainees...
during the study to maintain blinding of the participants towards the tools and avoid training to the test. Rather, the results informed the content that needed to be covered during each coaching session. In an implementation setting though, it would be useful to share the details of the objective analyses with the trainee as a transparency of results may aid learners in understanding deficiencies and help deal with negative emotions related to negative feedback.

Current in-training assessments used in competency based programs such as the DOPS and PBA (Intercollegiate Surgical Curriculum Programme, 2013; Intercollegiate Surgical Curriculum Programme, 2006-2014) combine global assessments with several task-specific items, yet many items remain general. The value of these evaluations is augmented by feedback comments made as free-text entries. The assessments though, lack detailed technical analysis as would be obtained through error analyses. Since assessments build the basis for meaningful feedback and serve as a reference guide, against which the trainee can learn to orientate self-assessments, it is of highest importance that routine continuous assessments become common practice in surgery. Still it is the lack of direct observations with the purpose of meaningful assessment that has been identified as a major barrier to competency-based training programs (Holmboe, 2014). Further risks to objective, fair feedback are the various sources of bias that can influence how assessments are made. For example, in workplace-based assessments it has been shown that few low scores were given by supervisors (Eardley, et al., 2013). This finding may be due to either trainees deferring assessments until late in training so that higher scores can be achieved (Eardley, et al., 2013), or may be due to the fact that supervisors may want to avoid confrontation and thus give higher grades in a face-to-face settings (Colletti, 2000). In addition, there is a high risk of bias, as the assessor is familiar with the trainee that is to be evaluated. For example, personal sympathies and positive views on specific performance aspects may unduly overshadow actual performance (halo bias), as would antipathy perhaps lead to unwarranted negative ratings (Thorndike, 1920). In addition, the observer may also be prone to recency bias especially when evaluating longer cases. This recency bias may cause the assessor to focus the evaluation on only the last or most memorable portion of the observation. Both rater biases potentially limit objectivity in direct-assessment settings. In addition, errors committed during operative cases have been identified as starting points for teaching interactions during operative procedures (Roberts, et al., 2012). This finding would suggest that error analysis should be also included in in-training assessments, as additional learning points may arise from focusing trainee
attention towards identifying technical errors. Currently though, workplace-based assessments are the only routinely performed structured assessments of operative performance and thus represent a realistic approach to routine evaluations. Despite the limitations due to various sources of bias and a lack of granularity for feedback, they may in future serve as the basis for post procedure debriefing and coaching.

An alternative method of assessment, which may reduce the risks associated with the biases previously described and may decrease the likelihood of grade inflation with subsequent loss of formative value of assessments, could be the supplementation of in-training assessments with video-review of performance. Video review has been described as a powerful adjunct to in-training assessment (Beard, 2008). The analysis of video material can be conducted in a standardized fashion using frameworks that have sufficient validity evidence to support the interpretation of results. Objective examples could be edited from the video-material to be used as video-feedback that may aid the trainer in structuring feedback sessions. The objective video material could further assist the learner in calibrating internal performance standards when referenced against exemplary behaviours using educational video clips. The training approach in this thesis detailed as comprehensive surgical coaching relied on video analysis as a means of generating objective valid feedback for coaching sessions. Video playback was used to aid participant’s self-debriefing processes, negating any negative sentiments that may be associated with perceived lower performance as the video review enabled the trainee to judge for themselves how they performed against examples of optimal performance. This process of assessment, feedback and debriefing led to significant improvements in skills and resulted in more accurate self-assessment abilities. Besides documenting skill progression, the regular assessments thus influenced training without requiring additional technical training exercises.

Comprehensive assessment represents the cornerstone of meaningful feedback and is an essential step required to structure the experiential learning experience in the OR. The results from our study on surgical coaching support the use of a comprehensive objective assessment methodology that is based on video review covering global, task-specific aspects of skill, as well as error modes and injury mechanisms.
7.3 Evidence-based Intra-operative Teaching - Structuring the Learning Experience and Avoiding Pure Discovery Learning

The OR as a learning venue has a long standing tradition. Despite the introduction and wide acceptance of surgical simulation as a means of educating future surgeons, the OR will remain central to resident training (Canter, 2011; Levinson, et al., 2010; Monkhouse, 2010). Effectiveness of simulation technology and curricula has been demonstrated to support the role of simulation in surgical education (Zendejas, et al., 2013). In the same vein, training in the OR needs to be scrutinized regarding effectiveness and appropriateness of teaching methods (Levinson, et al., 2010).

Several effective teaching behaviours for use in the OR have been identified including: structuring the educational experience using preoperative briefing techniques; intraoperative explanatory and questioning communication behaviours; postoperative debriefing; and formative feedback based on overall assessment of performance (Grantcharov, et al., 2007; Hamad, et al., 2007; Roberts, et al., 2009). In addition, training models based on principles of graded responsibility have been advocated including task deconstruction, and graded autonomy approaches such as in the Zwisch model (DaRosa, et al., 2013; Marangoni, et al., 2012). Nevertheless, deliberate teaching behaviours are inconsistently exhibited during operative encounters (M. Ahmed, et al., 2013; Scallon, et al., 1992). Potential learning experiences may also be missed through failed communication interactions (Roberts, et al., 2012). Furthermore, divergence in trainer-trainee perceptions of frequencies of OR teaching behaviours indicate that intraoperative teaching is not as clearly structured and obvious as would be expected (Butvidas, et al., 2011; Jensen, et al., 2012; Sender Liberman, et al., 2005). In the absence of deliberate instructional approaches and clearly defined learning goals, the OR becomes a discovery learning environment. In discovery learning, the learner will need to identify the pertinent content by themselves and integrate this new knowledge into their existing knowledge-base. This unguided process of pure discovery learning has been shown to be ineffective (Mayer, 2004), and thus may represent a missed learning opportunity.
The coaching intervention detailed in Chapter 6 draws on established instructional approaches, structuring the learning experience even after the encounter has occurred. During the coaching, the coach and trainee reviewed an experience using debriefing and feedback techniques, identified future learning goals, and sought out solutions to problems in technical execution. The operative encounter could be revisited through video playback which was used to generate dedicated learning points. For example, observed errors were used as a starting point for discussion on aspects of surgical technique, error avoidance, and rescue mechanisms. Subsequently, evidence-based video examples of relevant technical errors and adverse events were shared to emphasize relevance of specific maneuvers. This comprehensive approach combining the most relevant instructional strategies may compensate for a lack of intraoperative structure since the educational framework for future encounters is developed outside the OR through interactions between the coach and the trainee. The trainee can subsequently use this predefined structure even in an unguided environment. Similarly, the coaching intervention also aimed to empower the trainee to become a more proactive learner, seeking educational encounters during all operations. Through the use of briefing techniques, the trainee was encouraged to engage in communication before and after cases to determine learning objectives and seek out constructive educational communication behaviours by asking questions.

Although ideally structured training should be widely implemented and become common practice, with the goal of ensuring uniform educational quality, this may remain challenging considering all influencing factors. Optimally, the supervising surgeon would take on the role of the coach for the duration of the resident’s rotation. However, as was shown in Chapter 6, an external coach unrelated to the training environment can also ensure a productive learning experience without any modifications of training as it occurs within the actual OR. The additional structure in training through the use of coaching strategies led to significant improvements in skill, whilst trainees undergoing training without this extra structure were exposed to a heterogeneous learning environment. The remarkable finding that prior operative exposure in advanced laparoscopy, months in MIS rotations and training level did not predict BL surgical performance in LRYGB underscores the inconsistent learning that may occur in a largely unguided discovery learning context. After dedicated training using coaching though, case volume did predict performance, suggesting that structuring training is essential to ensure consistent skill acquisition. It must be acknowledged that all learners are different and that some
individuals may be naturally talented, requiring little guidance and practice, whilst others may require a clearer structure in training with numerous repetitions. The latter group may not substantially improve in skill in an unguided environment despite numerous months spent on designated rotations. To ensure that all trainees acquire the minimum skills and achieve competency it is therefore necessary to structure training in the OR and detect individual deficiencies of performance.

Furthermore, content of training needs to be standardized. As numerous different faculty members will be involved in the training of surgical residents, opinions on technical task execution will likely vary. Much of this variation is beneficial for the trainee, as they may construct their knowledge based on multiple sources, yet the influences of the “hidden curriculum” cannot be ignored. Surgeons need to be aware of their role model function, as trainees will adopt attitudes and performance traits that they have observed during their workplace-based training. The Delphi surveys described in Chapter 4 highlighted that, although expert surgeons from differing backgrounds shared a common conception of appropriate techniques in laparoscopic surgery, variability in views was also noted. Essential definitions of principles of what is considered correct and incorrect are required, and these should be shared with trainees. These definitions should ideally be evidenced-based, and widely accepted. Conceptions of acceptable and unacceptable though vary between surgeons and as the staff surgeon represents a role model for trainees, the particular techniques demonstrated will influence how trainees acquire skill over time. In trained surgeons, variations in task execution may have little clinical consequence as they compensate for any errors with high skill levels and sound understanding of rescue mechanisms. Trainees though may not be able to compensate and thus should be trained to adhere to fundamental principles including safe needle handling practices, appropriate visualization of tasks and anatomy, and tissue handling principles. The consensus list generated through the Delphi surveys described in Chapter 4, for example, could be used by program directors as a means of ensuring that trainees understand which principles are essential in laparoscopic surgery. The list could also be shared with the trainers, to make them made aware of basic conceptions on which a number of surgical experts have agreed. Program directors may also opt to supplement these lists with the opinions of the program’s faculty to focus the basic requirements to the appropriate setting, yet ensuring that all educators share a common understanding of the technical principles to be taught throughout the program.
The teaching behaviour of using narrative content during cases to raise awareness has been reported as a useful instructional approach in surgery (Hu, Peyre, Arriaga, Roth, et al., 2012). Although memorable, some of these “war stories” may be fictional and based on urban legend and thus bear little resemblance to evidence-based education (Hu, Peyre, Arriaga, Roth, et al., 2012). Nevertheless, these stories are believed to enhance retention of the learning point made and are considered common in the operative environment (Hu, Peyre, Arriaga, Roth, et al., 2012). The principle of storytelling has a long history and has frequently been used in training settings outside the health-care environment (Spooner, 1993). Spooner (1993) described the most important features of storytelling to include the fact that the listener should be able to identify with the story, and that the story should be concrete, credible and remarkable. In medicine, storytelling was found to focus around topics of personal training experiences, practice changes, and near misses or adverse events (Hu, Peyre, Arriaga, Roth, et al., 2012). The clinical teaching points included issues of surgical technique and management, decision making, and error identification (Hu, Peyre, Arriaga, Roth, et al., 2012). Despite often deemed useful, these stories are not evidence-based and could be supplemented with real evidenced examples. Such examples could be generated through video review of routine procedures or could be derived from root-cause analyses of adverse events. In Chapter 5, an analysis of selected identified intraoperative minor events revealed that even in successful surgeries numerous examples for learning experiences can be obtained. Similarly, the knowledge gained through the retrospective video review of routine cases as well as cases deemed to have resulted in an adverse situation, could deliver the evidence base for future stories that could be told with educational intent.

In addition, video review could also deliver an evidence-base to guide curriculum content in technical skills training. Through the review of routine cases performed by trainees, a general impression of learning needs can be generated. For example, we screened the coaching content from our study that was based on video reviews, to identify common learning themes. Surprisingly, general topics such as “use of camera”, “working bimanually”, “keeping loaded needle in view”, and “loading needle techniques” were common topics. These themes suggest that basic simulation curricula and didactic training content in surgical curricula could focus more on elemental principles in laparoscopic surgery. Therefore, using video review may represent a very useful tool for educators when planning future training curricula, and as an evidence-based means of guiding training program specific content.
7.4 Self-assessment

Self-reflection is an important process in personal development as it represents a core step in experiential learning. Still trainees have been shown to prefer external assessment to self-assessment, and self-assessments have been perceived as being stressful (Evans, McKenna, & Oliver, 2001). Furthermore, although self-assessment is central to self-regulation, a process necessary for continuing professional development (Epstein, et al., 2008; Silver, et al., 2008), physicians have repeatedly been shown to possess poor self-assessment skills (D. A. Davis, et al., 2006). This lack of self-assessment ability is of concern since external feedback becomes rarer as medical professionals progress in their career, with senior doctors rarely receiving peer input (Evans, McKenna, & Oliver, 2002). Furthermore, Epstein et al. (2008) detailed that physicians as self-assessors rely on the integration of external data to guide their assessments, yet meaningful feedback and objective data is frequently lacking. For this reason, assessments are frequently made based on internal data that is influenced by the assessor’s self-image (Epstein, et al., 2008). As a result, teaching self-assessment skills and ensuring appropriate measures against which the assessor may reference the self-evaluations is of high relevance. Several researchers have focused on topics pertaining to trainability of self-assessment skills especially outside health-care. In child education, for example, evidence remains mixed with studies supporting an effectiveness of instruction in self-assessment in math achievement (Ross, Hogaboam-Gray, & Rolheiser, 2002). Other studies showed no effect, for instance on music performance skills, or self-assessment accuracy (Hewitt, 2011). Although this topic is of particular relevance to the self-regulated health-care professions, the subject of teaching self-assessment skills in medical school or specialty training remains under-investigated. On the question of training self-assessment, Eva and Regehr (2005), suggested that physicians may best benefit from identifying sources of external reference and using these in reflective processes as a means of self-improvement rather than attempting to increase self-assessment accuracy (Eva & Regehr, 2005). It is this openness to feedback with a willingness to accept critical input that may drive self-directed learning and professional development.

In the study reported in Chapter 6 structured coaching involving facilitated debriefing and feedback, achieved that participants’ self-assessments aligned with ratings by blinded observers,
while self-assessment of conventionally trained residents showed no correlations with expert ratings. Examination of the absolute differences between self-assessment scores and external ratings revealed that the score deviations in the conventionally trained group were approximately three times as large as those observed in the coached group. Further, the conventionally trained group’s scores showed substantial variability regarding the extent of deviation (data not shown). We hypothesized that the coaching intervention would induce a greater understanding of the general principles of LRYGB that could serve as an external reference source for self-assessment. This understanding of correct and incorrect was fostered through active debriefing, video-review and viewing of exemplary behaviours. We theorized that subsequent self-assessments would be guided by this external reference, facilitating more accurate self-assessments when using criterion-referenced rating tools. In our study, neither group had been familiarized with the rating tools or scales before the self-assessment exercise. Therefore, exact calibration to the anchors, as was expected of the trained raters, was not anticipated. Since both groups used the same criterion-referenced, anchored tools, we expected that both groups would be able to interpret their performance in the context of these scales, showing a correlation between their scores with external measures (expert ratings), even if the single measures individually deviated from the “gold standard” of the external rating. This correlation though was not observed in the conventionally trained group, suggesting that they lacked a reference framework against which they could judge their performance, despite using explicit assessment tools.

Ward et al. (2002) criticized methodological aspects of reporting self-assessment accuracy in the medical literature suggesting that causes for the presumed inaccuracy of self-assessments may be sought elsewhere. Causes for inaccurate self-assessments postulated by the authors were potential invalidity of external reference criteria (gold standard) against which the self-assessments were compared, use of ambiguous scales in self-assessment that may hinder the self-assessor referencing abilities, and concerns with statistical methods used to determine self-assessment accuracy (Ward, Gruppen, & Regehr, 2002). We can refute these concerns in our study as the external criterion against which the self-assessments were measured was derived from the averaged measures of two independent raters, with both raters showing excellent inter-rater reliability. In addition, explicit scales were used during self-assessment, and although only correlation statistics are shown in detail in Chapter 6, additional data analyses revealed marked
absolute deviations from the external ratings in the conventional group as compared to the coached group’s measurements. We thus believe, that the coaching intervention aided the participants in developing a mental reference guide against which they could measure their performance, and that they were open to feedback and critical review, as this was a common feature throughout the training. Consequently, structuring training to incorporate the approaches detailed in the coaching intervention may also be a step towards teaching self-reflection and openness to external input thus improving self-assessment abilities.
7.5 Considerations for Technical Skills Training in Postgraduate Training

In modern training programs, surgical education has been augmented through simulation curricula to improve surgical skill outside the OR. These simulation-based curricula have been shown to deliver superior training results when compared to no intervention (Zendesjas, et al., 2013). Still it must be acknowledged that simulation-based curricula, besides offering a safe venue for practice, often also incorporate established instructional approaches such as feedback and deliberate and distributed practice protocols to structure training (Zendesjas, et al., 2013). As a result it is not surprising that simulation leads to superior skill acquisition. Since it has also been recognized that the OR will remain a central venue for learning in surgery, simulation in surgical training should be aimed at advancing trainees in their early learning curve so that these can enter the OR as “pretrained novices” (Gallagher et al., 2005). Simulation must, therefore, be seen as a complementary training method, rather than an alternative to training in the OR. Limitations to implementing simulation training have been recognized predominantly as factors related to resident motivation. Motivation may be influenced by fatigue, availability of protected time, total work-hours, clinical responsibilities, and fidelity and realism of models (Stefanidis & Heniford, 2009). Furthermore, residents do not believe that simulation is a good substitute for practice in the real OR (Shetty, Zevin, Grantcharov, Roberts, & Duffy, 2014). Thus, they may be less motivated to attend, if this resulted in a potential loss of OR time (Stefanidis & Heniford, 2009). Further factors that may limit the comprehensive integration of simulation in surgical training may be the need for adequate resources including funding, space, equipment, faculty, and facility personnel (Stefanidis & Heniford, 2009). In addition, a recent survey of surgical residents revealed that with progressing training experience, residents’ preferences in simulation modalities also shift (Shetty, et al., 2014). Junior residents value box-trainers and VR simulators, while advanced trainees prefer live animal models and explanted tissues over VR simulators (Shetty, et al., 2014). This finding may be explained by the training focus at different stages of residency. Junior residents will need to acquire the basic skills required of a surgeon, which can easily be achieved through deliberate practice in simulations until proficiency in these skills has been obtained. Senior residents though need to integrate knowledge and skills to manage more
complex procedures, as well as to manage intraoperative complications (Gallagher, et al., 2005). Simulating these settings with adequate realism may be challenging. These limitations need to be acknowledged, and transition to training in the OR as “pretrained novices” needs to be transparent to trainees so that they understand the benefits of both training modalities. Furthermore, much can be learned from the successes of simulation as evidenced by educational research. Studies evaluating the effectiveness of simulation have commonly included additional instructional approaches, a phenomenon that is still lacking in the OR.

The coaching intervention reported in Chapter 6 pairs well with simulation training even though this was not explicitly tested. For example, many identified learning goals represented fundamental steps or tasks that could have also been addressed in deliberate practice sessions in a simulated environment. Training can be tailored by exploring coaching profiles. Individualization of simulation training through determining learning needs may also avoid unnecessary repetitions of tasks in which the trainee is already competent. Tailored simulation training may, therefore, reduce time requirements. This focused approach may aid in motivating senior trainees to return to the simulation facility for practice sessions. Furthermore, deliberate practice embedded in coaching programs could likely be performed without direct faculty supervision, as training success would be evidenced through improved performance in subsequent operative cases. These operative cases would be the topic of subsequent coaching sessions where feedback could be afforded to the trainee, indirectly reflecting back on the deliberate practice session.

In contrast to simulation though, the benefits of coaching as in our context as a standalone approach included that residents were easily motivated to participate as the training was brief and did not take time away from the OR or other clinical duties. In addition, no specific facilities or equipment were required and therefore, this teaching approach could also be implemented in settings where residents do not have access to a simulation facility. In summary, coaching as a method of structuring experiential learning in the OR needs to be viewed as complementary to simulation, with both modalities best combined for maximum effectiveness.
7.6 Conclusion

Surgical education in the OR remains essential in modern competency-based training. The historical training approaches that involved countless hours of direct observation and practice in the OR may no longer be applicable. Modern health-care emphasizes the need for patient-centered, safe, time- and cost efficient surgery. As educational encounters become fewer, teaching approaches in and around the OR also need to change. Teaching must become purposeful, focused on capturing all potential learning experiences and aimed at guiding the trainee towards becoming a self-directed learner. Although numerous instructional approaches have been described that may be beneficial to learning in the unguided OR environment, the majority of evidence suggests that these teaching behaviours remain rare. As a consequence, surgical programs must devise structured approaches to training in the OR. Similar to other curricular components integrated in residency such as weekly lectures, seminars, rounds and simulation sessions, so should structured training in the OR become a formal curricular teaching priority. The comprehensive surgical coaching technique detailed in this thesis is such a structured approach that could aid in formalizing educational encounters of everyday operative experiences. This coaching combines relevant instructional methods that have individually been shown to be effective in workplace-based or situated learning environments. Modern concepts, such as comprehensive objective assessment including error analysis, have also been incorporated in the coaching approach. Comprehensive assessment, besides informing learning needs, may also be used as a means to document competency-acquisition throughout training. Furthermore, error analysis can be used to teach principles of error recognition, awareness, and mitigation, which addresses educational content not formally taught within current residencies. Lastly, this type of training may aid in structuring the discovery learning experience, even after the experience has already occurred. Thus, even when intraoperative teaching lacks formal structure, coaching can be used to guide trainees through reviewing their experience, capitalizing on the educational value of each operative case completed. Beyond technical skill improvement through coaching, the resident as a learner may also become empowered to actively seek out educational encounters, as coaching aims to define personal goals. This goal-driven approach may thus influence intraoperative behaviours as the resident may be stimulated to ask questions, state personal learning goals, and seek debriefing or feedback from the respective educator in the
OR. Familiarizing the learner with concepts of critical self-reflection also improves self-assessment abilities, a skill that is paramount in self-directed life-long learning.

In summary, surgical coaching is a versatile approach to structuring the experiential learning experience in the OR which significantly improves surgical technical skill and decreases technical errors.
This chapter summarizes the relevant limitations of the research projects detailed in this thesis.

Although the studies detailed in this thesis add significantly to the existing knowledge base and supply new evidence in support of evidence-based training content and instructional approaches, limitations of this research also need to be acknowledged. Specific limitations of each study are included in each chapter at the end of the respective discussion sections. In the following section, I will, for this reason, discuss the general limitations that apply to all studies conducted within this thesis.
8.1 Lack of patient outcome data

In Chapter 3 I have described the design of an assessment framework to determine technical errors in laparoscopic surgery, specifically in LRYGB. The framework was based on existing established concepts derived from other high-risk industries and modified for health-care as a means to explore human error. From these frameworks, the GERT framework was reduced to include only error modes that represent the smallest unit of incorrect task execution. As detailed in Figure 4, these error modes could remain entirely without consequence or lead to an event that represents the next level of the hierarchy. The role of errors causing events could be documented in the study described in Chapter 5, but the relevance of all errors that do not cause an event remains unclear. Potentially the sum of these errors may affect the overall course of an operation through influencing procedure duration, or could influence outcome through summative effects that are yet to be determined. The relationship between error count and procedure duration was found to be significant, in that correlations between the two measures were identified (Bonrath, Zevin, et al., 2013). Whether more errors resulted in prolonged procedure duration through extra moves or whether longer procedures gave rise to more opportunities to make errors remains unclear. How and if these interactions play a role in patient outcome cannot be answered by the studies reported in this thesis. This question will need to be answered through subsequent research in future prospective studies. Similarly, although the events identified in Chapter 3 and Chapter 5 represented situations that had the potential to cause harm, all were identified and rectified at the time of surgery. These events could be termed “near misses” as in other situations harm may have resulted. It should be noted, however, that the studies cannot provide the evidence base to document how these events could potentially lead to adverse outcome. Following the event classification proposed by Champion et al (2008), all events would therefore correspond to the group termed “minor events”. “Minor events” describe events that potentially require additional actions but failed to affect the overall course of the operation (Champion, et al., 2008). Nevertheless, reviewing performance based on all aspects of skill is of high relevance for training even if a direct causality can, at this point, not be shown. As a large-scale review of surgical performance at “staff level” showed that higher skill level corresponded to better patient outcome using global skill metrics, improving all aspects of skill seems essential (Birkmeyer, et al., 2013).
Similarly, the impact of the coaching training intervention on patient outcome was not investigated in the study reported in Chapter 6. Although it has been documented that resident participation in operative procedures may result in prolonged procedure durations, (S. S. Davis, Jr., et al., 2013; Hernandez-Irizarry, et al., 2012; Tseng, et al., 2011) evidence regarding morbidity remains mixed. In contrast, the majority of data on mortality rates suggest lower rates due to resident involvement (S. S. Davis, Jr., et al., 2013; Tseng, et al., 2011), likely as a consequence of higher rate of rescue in the case of complications (Castleberry, et al., 2013). Reduced morbidity rate associated with targeted training has to date only been shown in one study using simulation-based training to specific competency benchmarks (Zendejas et al., 2011). Zendejas et al (2011), reviewed postoperative complications including urinary retention and seroma, as well as length of hospital stay following laparoscopic inguinal hernia repair as outcome metrics in a randomized trial exploring effectiveness of simulation-based mastery learning (Zendejas, et al., 2011). These outcome metrics were appropriate in the setting of the study detailed by Zendejas et al. (2011), as laparoscopic inguinal hernia repair is commonly performed as a day case, and complications such as urinary retention are common well known complications. In LRYGB, on the other hand, surgery and treatment protocols are more complex. The procedure component performed by the residents in our study was only a portion of the overall procedure, and complications may arise from all substeps of the operation. Furthermore, our study was conducted in an educational environment under supervision. Therefore, any poor resident performance would have been recovered by the attending staff in all cases. Furthermore, postoperative management following LRYGB includes an in-patient stay of two to three days, a period during which care is delivered by a group of professionals rather than an individual. In addition, concurrent medical conditions are common in the bariatric population, which may also influence convalescence after surgery. Therefore, measuring the effectiveness of the training intervention through markers such as complications or length of stay as detailed by Zendejas et al. (2011) does not appear reasonable in our setting. This opinion was also shared by Cook and West (2013), who criticized the recent focus on “outcomes research” (Cook & West, 2013). The authors make several compelling arguments against reporting outcomes as a measure of educational effectiveness including considerations of effect dilution in the complex treatment environment, feasibility of studies that require a large sample sizes to show differences in frequencies of rare complications, and the fact that surgical education needs to focus on the “learner” as the end-user of the training intervention (Cook & West, 2013). Therefore, in our
study design, we chose outcome parameters that would enable an unbiased assessment of the performance of only the resident, limiting influences of other staff involved in the care of the patients, to document effectiveness in teaching the skills required in LRYGB.

Furthermore, as all data collection was anonymous with regards to patients, no information was available concerning the complexity of the case. This limitation needs to be acknowledged as more complex cases may predispose the participant to making more errors or exhibiting lower skill. To control for the factor “complex patient”, trainees were randomized. Consequently, effects would likely have influenced both groups. Furthermore, in a training environment, the supervisor decides which case is appropriate for which trainee following principles of graded responsibility. Therefore, it is unlikely that a trainee would have been required to complete a complicated case unless the supervisor felt that the trainee was advanced enough to handle the case. Therefore, it is unlikely that this factor played a relevant role in the comprehensive coaching study.

8.2 Generalizability

The studies reported in this thesis were all conducted within one educational environment. Therefore, the results will need to be viewed in the context in which they were gathered.

In Chapter 3, the GERT framework was tested using one complex procedure using videos from one educational archive. Good reliability was seen for total errors and events (as the primary outcome), and several sources of validity supported the interpretation of the results that the measures reflect an aspect of surgical performance. Still, not all task groups achieved acceptable levels of reliability as some error observations were infrequent. As the GERT is used to measure error in other procedures and other disciplines more evidence to support the individual task groups will likely be gathered since the spectrum of tasks vary in different cases and specialties. Therefore, future studies will be needed to supplement the existing validity evidence until all aspects of the framework have sufficiently been explored. This process of generalization will also show whether relevant task categories are currently missing from the framework (as
measured through entries in the category “others”), or identify items that are redundant. This data may lead to future modifications of the tool. For our purposes though, the GERT was used in the context in which it was tested, and only total scores were used as final outcome metrics since these were supported by sufficient validity evidence in the design study.

In addition, reliability of the GERT clearly depends on the training of the raters using the tool, as a lack of understanding regarding includable maneuvers threatens reliability of measurements. Raters need a conceptual understanding of the discreet differences between errors as largely inadvertent maneuvers and intentional skill actions in order to avoid a redundant overlap between rating frameworks. During the design of the GERT, we therefore felt it was necessary to supplement the rating structure with a guide to erroneous maneuvers. This guide was obtained through the international Delphi process detailed in Chapter 4, which generated a list of items that were widely considered relevant in a training setting. This guide list is generalizable as it was drawn from the opinions of educators and surgeons from a wide geographic region and varying surgical backgrounds. We believe, that in training raters to use GERT these items should be discussed to clarify includable errors observed. Each application of the GERT should also be supplemented with a set of “rules” that are to be applied to rating as local variations in conceptions of surgical task performances may vary. Detailing what rules were applied to rating will much facilitate comparison of results from differing research or training settings.

The results described in Chapter 5 reflect the detailed analysis of error-event patterns in LRYGB in one educational environment, although the original videos had been obtained from multiple different surgeons and sites within this environment. The results, therefore, paint a picture of common patterns in the very specific educational setting. Depending on practice variations, different patterns may emerge in other settings. For example, in our setting the specific anastomosis technique of using a circular stapler was not represented. Consequently, typical events related to use of the stapler would not have been observed. Thus, we suggest that educators use the technique of targeted video review of routine procedures as a means to explore the educational environment in which they work. The information obtained may readily guide future educational content. For example, recurrent injury mechanisms and frequent errors may be the basis of simulation curricula or gaps in knowledge may be addressed in lectures and seminars.
Lastly, the limitation of generalizability of the coaching intervention needs discussing. In this study, we used a pragmatic trial design, yet we controlled the factor of the coach to standardize training delivery. One coach delivered all interventions; as a result, the role of the coach remains central to the effectiveness of the intervention. It is possible that this effectiveness may be lost depending on the personal attributes of the coach and whether or not they can build a rapport with the coachees. Using a coaching structure such as the PRACTICE or GROW frameworks may prove beneficial to limit this influence, as coaching structures offer new coaches guidance throughout the process (Palmer, 2007). Moreover, to enhance reproducibility in our study, coaching was also supplemented with instructional approaches that have been shown to be effective, and that can be replicated in other settings (feedback, debriefing, behaviour modeling). Another limitation of our study design may be the particular learning environment. In other environments, different factors may be present that could hinder or enhance the integration of new knowledge gained through coaching for example via the hidden curriculum. Therefore, in order to implement coaching successfully, the educational context and potential barriers need to be identified so that instructions can be modified to facilitate the integration of the training component as needed. In addition, the model we chose was a single coach model. Effects may differ if coaching is delivered by several coaches (for example in the setting where coaching is alternated between faculty members). Therefore, any implementation of a novel coaching program (as in any new educational intervention) will need to undergo evaluation to determine whether the effectiveness is maintained outside the research setting as a means of justifying additional teaching resources and program time allocation.
This chapter highlights the possibilities for future research based on topics explored in the prior chapters.

The studies reported in this thesis supplement the existing knowledge base, and the new evidence gained could serve as the basis for future research. In addition, future projects, primarily targeting implementation and generalizability of the coaching intervention, could be conducted to address some of the limitations stated earlier.
9.1 Generic Error Rating Tool – Error Analysis

Future work on the GERT tool could include exploring generalizability as a generic tool. For this purpose, analyses of other general surgical procedures as well as laparoscopic procedures from other disciplines such as gynaecology or urology could be conducted. Through the analysis of different procedures, additional validity evidence can be gathered for errors in task groups that were infrequently observed in LRYGB.

Besides pairing the analyses of the GERT with other performance metrics derived from observational tools such as the OSATS (Martin, et al., 1997), error frequencies need to be explored as predictors of patient outcome and operative costs (as a consequence of duration and additional resource requirement).

Another theory that could be studied in future research is that errors may not only be a result of a lack of skill but could also reflect procedure complexity. This theory is based on the observation from the GERT design study that showed that skillful surgeons still committed errors, despite scoring at the top end of the OSATS scale. Since these surgeons were deemed to be very skillful, other factors, potentially related to the complexity of the step, may have contributed to the higher error count. During the assessment of procedures in the design study of GERT we noticed that errors did not occur at a linear fixed rate, rather they could be seen to cluster in specific steps. For example, the introduction and set-up phases of a procedure were less error prone than the steps related to the GJ creation that incurred the most errors even when skilled surgeons completed the step. This knowledge could be used to explore the step complexity of procedures as a means of identifying the most error prone steps, warranting dedicated training. For this purpose, the performance analysis of procedures by highly trained surgeons beyond the learning curve would be required. Ideally variations in procedure execution (minor technical variations) could also be explored to detect procedure variations that are the least error-prone. This information could serve as a basis to standardize the training of these operations in postgraduate surgical education.
Error analysis could also be used as an educational tool in itself. Considering the findings that understanding and recognizing errors in end-product analysis is related to surgical skill and knowledge (Bann, et al., 2003; Bann, et al., 2005), error analysis of brief video clips may serve as a testing tool to assess residents’ error awareness and understanding. The residents’ level of knowledge could be objectively measured by using exemplary clips of erroneous performance. This cognitive test of surgical performance could complement other knowledge tests in competency assessments.
9.2 Error- Event Analysis

Future studies could explore the implications and relevance of error and event distributions in operative procedures.

The error-event analysis conducted in Chapter 5 involved the secondary review of previously identified events. An additional finding in this study was that a large proportion of events identified related to steps of dissecting the lesser curve of the stomach during gastric pouch creation. This finding could be a consequence of the sampling technique used in the study (convenience sample) or could reflect a more hazardous step within the procedure. Future error-event analyses should focus on the assessment of error distribution and event ratios in single surgeon cases. Using error to event ratios, steps could be identified where errors are at a higher risk of being consequential. As Cuschieri (2003) detailed, the step during which errors occur may have more relevance regarding patient outcome, than the error itself. The concept of “hazard zones” refers to steps of procedures where an error will carry a greater consequence (Cuschieri, 2003). As a result understanding hazard zones in routine procedures is of paramount to improving patient safety.

By understanding consequences and risks associated with errors, possible weighted rating versions of the GERT could be developed. Although each error is of importance when reviewing surgical performance in training and feedback, the weighting of scales to reflect severity or consequence may be useful for assessments aiming at predicting patient outcomes. Before any weighting of errors can be done though, a full understanding of errors and their relationships with events and outcomes is required. In addition, as errors can never be fully avoided, studies could focus on determining normal distributions of errors in routine cases, defining limits of acceptable error rates and identifying the cases that lie beyond the expected norm and that should be more carefully examined.
9.3 Surgical Coaching

The coaching study described in Chapter 6 opens numerous avenues for future research. Especially topics of implementation of coaching into routine residency training, relevance of distance coaching, and coaching of trained surgeons are stimulating research fields.

As discussed in Chapter 6, factors that will influence the implementation of coaching are the targeted training group (junior, senior residents or staff), the type of coaching frequency selected (continuous program versus interval coaching), the type of procedure to be coached, and who should take on the role of the coach. During the coaching intervention, we observed that trainees became more comfortable with coaching as the rotations progressed, with sessions becoming shorter and also less directive. This trend would suggest that trainees became less dependent on the coach and more self-directed. A resident that is self-directed with regards to identifying learning needs may continue the process on his/ her own even after the coaching sessions have ended. To test this hypothesis, retention of skill and review of self-debriefing skills would be required at set time intervals after conclusion of a coaching intervention. This type of study would aid in determining whether coaching in surgery needs to be a continuous feature or whether interval refreshers suffice to sustain the effects. Further, the impact of coaching on self-assessment could be monitored as a marker for when coaching interventions may be useful. Through regular video-based self-assessments during residency, educators may identify the time point at which corrective external feedback and facilitated debriefing may be useful to ensure that learners remain orientated to their performance. Since life-long learning is paramount in medicine, promoting the skills needed to determine learning goals is essential in residency and, therefore, future research should focus on the use of coaching to improve self-assessment abilities.

Future research could also be directed at determining who should take on the role of the coach. Different instructors as coaches including the supervising staff (for example as portion of in-training assessment), or designate educators (external to the training environment), or a hybrid of both could be explored. Coaching effectiveness in a distance setting using tele-mentoring could
also be the subject of future research as effectiveness of coaching may be lost in a lack of direct personal interaction.

In the setting that coaching is implemented into routine residency programs, research should aim to identify any potential barriers to coaching as well as organizational factors that may enhance the effectiveness of coaching such as coordinating protected time to receive and deliver coaching interventions. Program evaluation of implementation would aim to review factors such as attendance, time requirements, costs, perceived usefulness (both trainers and trainees), and performance metrics (video review) over the duration of participation in the program.

All of the research reported in this thesis focuses on surgical training in residency. Similar training concepts and self-evaluation could also be useful in trained surgeons. Subsequently, research could focus on coaching constructs applicable to practicing surgeons, possibly as a variant of distance learning using tele-mentoring.


Holmboe, E. S. (2014). Realizing the Promise of Competency-Based Medical Education. *Academic Medicine : Journal of the Association of American Medical Colleges. doi: 10.1097/ACM.0000000000000515*


Intercollegiate Surgical Curriculum Programme. (2013). Guidance for using Surgical DOPS

Intercollegiate Surgical Curriculum Programme (ISCP). (2006-2014). Overview of the
Assessment System, from https://www.iscp.ac.uk/surgical/assessment_overview.aspx

Intercollegiate Surgical Curriculum Programme (ISCP) (2006-2014). Procedure Based
Assessment (PBA) Retrieved 11.12.2014, from
https://www.iscp.ac.uk/surgical/assessment_pba.aspx

How to train surgical residents to perform laparoscopic Roux-en-Y gastric bypass safely.

Irby, D. M., & Wilkerson, L. (2003). Educational innovations in academic medicine and

Journal of Evidence Based Coaching and Mentoring, 6(2), 100-113.

teaching. Journal of Surgical Education, 65(6), 436-440. doi:
10.1016/j.jsurg.2008.07.006


procedure-based assessment is feasible in clinical practice. Annals of the Royal College of

assessment augments development of videoscopic suturing skill. Journal of the American

feedback in the operating room: a gap between resident and faculty perceptions.

human reliability analysis. Applied Ergonomics, 29(6), 409-414. doi:
S0003687098000167 [pii]

British Medical Journal, 311(7001), 376-380.


10.1016/j.surg.2010.01.011


### Appendix 1: Generic Error Rating Tool (GERT) Checklist from (Bonrath, Zevin, et al., 2013)

<table>
<thead>
<tr>
<th>Surgical task group</th>
<th>Error mode</th>
<th>Time of observation</th>
<th>Total number</th>
<th>Event (description/time)</th>
<th>Mechanism of event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal access</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of retractors</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of energy devices</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dissection and grasping</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cutting, stapling and transection</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clipping</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suturing</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Too little force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wrong orientation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of suction</td>
<td>Too much force/distance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inadequate visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other unclassified</td>
<td>Description/time:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Copyright Acknowledgements

Chapter 1, section 3.2.4: Copyright permission from Springer Science+Business Media


The final publication is available at: http://link.springer.com/article/10.1007%2Fs00464-013-2827-5#

Chapter 3: Copyright permission from John Wiley and Sons license number 3555540031112


© 2013 British Journal of Surgery Society Ltd. Published by John Wiley &Sons Ltd. British Journal of Surgery 2013; 100: 1080–1088

Original full-text article available at: http://onlinelibrary.wiley.com/doi/10.1002/bjs.9168/abstract;jsessionid=905F027BD3535FEB87764FD0F4EB39A0.f01t01

Chapter 4: Copyright permission from Springer Science+Business Media


**Chapter 5:** As per licence agreement with BMJ Publishing Group Limited and co-owner, the ownership of copyright remains with the authors. The copyright owners retain the rights for non-commercial use including the right to publish all or part of the accepted manuscript in a book essay, position paper, or other non-peer reviewed publication authored by the contributor.


**Chapter 6:** Wolters Kluwer Health Lippincott Williams & Wilkins grants the Authors permission to reuse their full-text article free of charge (non-commercial/ non-for-profit purposes) for the purpose of reuse in a dissertation.


**Figure 1:** Copyright permission from John Wiley and Sons license number 3555530231952


Table 1: Copyright permission from John Wiley and Sons license number 3555531189483
