

ASSESSING PERFORMANCE AND CONSTRUCT VALIDITY
OF LAPAROSCOPIC SURGICAL SIMULATORS

by

JOANNE LIM

B.Sc. (Eng), The University of Guelph, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Mechanical Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

December 2006

© Joanne Lim, 2006

Abstract

The objective of this work is to assess the construct and performance validity of two laparoscopic surgical simulators. Currently, the evaluation of surgeons is considered subjective and unreliable, and this is a reason why surgical educators have been studying surgical simulators as a method to quantitatively assess surgeons. But we must find out if these simulators are valid and reliable methods for training and assessing surgeons. We have designed an experimental surgical tool and data collection system to quantitatively measure surgeon motor behaviour in the operating room (OR). Our experimental system collects kinematics and force/torque data from sensors, and we have developed a sensor fusion algorithm to be able to extract high frequency and continuous kinematics data. We have collected data from surgical residents (PGY4), and compared it to expert surgeon data to investigate construct validity of both a physical simulator and virtual reality (VR) simulator. We also study the performance validity of both the simulators by comparing measurable quantities, such as force and kinematics, on the simulators with that collected in the OR. To examine differences in our contexts, we use the Kolmogorov-Smirnov statistic. According to our intrasubject intersetting (OR, VR, physical) comparisons, we see large differences between the OR and VR simulator, leading to the conclusion of poor performance validity. Conversely, we see smaller differences between the physical simulator and the OR, and therefore showing fair performance validity. In our interlevel (expert vs. resident) comparisons, we see that the VR simulator shows poor construct validity with little difference detected between skill levels, while the physical simulator seems to be able to detect differences in some performance measures and can be considered to show fair construct validity.

Table of Contents

Abstract.....	ii
Table of Contents.....	iii
List of Tables.....	viii
List of Figures.....	ix
Acknowledgements.....	xiii
Chapter 1: Introduction and Literature Review.....	1
1.1 Introduction and Objectives.....	1
1.2 Minimally Invasive Surgery.....	2
1.2.1 The Challenges of MIS for Surgeons.....	4
1.2.2 The Challenges of MIS for Surgical Educators.....	6
1.2.3 Reasons for Using Surgical Simulators.....	6
1.2.3.1 Surgeon Certification.....	6
1.2.3.2 Equipment Design and Evaluation.....	6
1.2.3.3 Transfer of Training.....	7
1.3 Current Training Methods.....	7
1.4 Current Methods of Surgical Performance Assessment.....	9
1.5 Simulator Validation.....	11
1.5.1 Construct, Performance and Concurrent Validity.....	13
1.6 Research Question.....	14
1.7 Developing a Quantitative Assessment Method.....	15
1.7.1 Kinematics.....	16
1.7.2 Forces/Torques.....	17
1.8 Project Goals.....	17
Chapter 2: The Hybrid Experimental Laparoscopic Surgical Tool for Performance Measure Assessment.....	20
2.1 Introduction.....	20
2.2 Laparoscopic Surgical Tool.....	21
2.3 Performance Measures.....	22
2.3.1 Kinematics.....	22

2.3.1.1 Optoelectronic Position Tracking.....	23
2.3.1.1.1 Other Kinematics Options.....	24
2.3.1.2 Electromagnetic Position Tracking.....	25
2.3.2 Force/Torque.....	27
2.3.3 Sensor Bracket Design.....	28
2.3.3.1 Force Balance.....	31
2.3.4 Grip Force.....	37
2.4 Force/Grip Data Processing.....	38
2.4.1 Grip Calibration.....	38
2.4.2 Gravity Effects Calibration.....	42
2.5 Kinematics Data Fusion.....	43
2.5.1 Data Fusion Introduction.....	43
2.5.2 General Data Fusion.....	43
2.5.2.1 Fusion Methods.....	45
2.5.4 Kinematics Data Fusion Technique.....	45
2.5.4.1 Data Fusion Technique Details.....	46
2.5.5 Error Analysis.....	54
2.5.5.1 Analysis Method.....	54
2.5.5.2 Results of Error Analysis.....	55
2.5.5.2.1 Computer Generated Data.....	55
2.5.5.2.2 Laboratory Data.....	57
2.5.5.2.3 Operating Room Data.....	59
2.5.6 Discussion of Kinematics Data Fusion.....	61
2.5.7 Conclusions for Kinematics Data Fusion.....	62
2.6 Discussion and Recommendations.....	62
2.6.1 Kinematics.....	63
2.6.2 Force.....	63
2.6.3 Recommendations.....	64
 Chapter 3: Experimental Methods for Assessing Validity of Laparoscopic Surgical Simulators.....	 66
3.1 Introduction and Objectives.....	66
3.2 Subjects and Settings.....	67
3.2.1 Settings.....	68
3.2.1.1 Operating Room.....	68
3.2.1.2 Virtual Reality Simulator.....	68
3.2.1.3 Physical Simulator.....	69
3.3 Performance Measures.....	70
3.3.1 Kinematics.....	70
3.3.2 Forces.....	70

3.4 Equipment Used.....	73
3.4.1 Video Data.....	73
3.4.2 System Component Integration.....	74
3.4.3 Data Acquisition Software.....	75
3.5 Data Collection.....	76
3.5.1 Operating Room Study.....	76
3.5.2 Simulator Data Collection.....	77
3.6 Data Post-Processing.....	78
3.6.1 Kinematics Data Registration and Calibration.....	79
3.6.2 Force/Torque Data Registration and Calibration.....	79
3.6.2.1 Force/Torque Data Registration.....	79
3.6.3 Raw Data Synchronization.....	80
3.7 Electrosurgery Unit.....	81
3.7.1 ESU Effects.....	82
3.7.1.1 Removal of ESU Effects.....	83
3.8 Task Comparisons.....	86
3.8.1 The Dissection Stage.....	86
3.8.2 Data Segmentation.....	87
3.8.2.1 Data Segmenting.....	87
3.9 Setting Comparisons.....	88
3.9.1 Kolmogorov-Smirnov Statistic.....	89
3.9.2 Comparisons.....	90
3.9.3 Assigning Confidence Intervals.....	90
3.9.4 Dependent Data and Moving Block Bootstrap.....	90
3.9.4.1 Measurement Resolution.....	91
3.10 Discussion.....	93
Chapter 4: Results of a Quantitative Study to Assess Laparoscopic Surgical Simulator Validity.....	95
4.1 Introduction.....	95
4.2 Results.....	95
4.2.1 Context Comparisons.....	96
4.2.1.1 Surgical Residents.....	96
4.2.1.2 Expert Surgeons.....	97
4.2.2 The D-Value.....	97
4.2.3 Presentation of Results.....	98
4.2.3.1 $\Delta 1$: Intrasubject Intraprocedural OR comparisons.....	98
4.2.3.1.1 Resident 1: Intrasubject Intraprocedural OR.....	99
4.2.3.1.2 Resident 2: Intrasubject Intraprocedural OR.....	101
4.2.3.1.3 Resident 3: Intrasubject Intraprocedural OR.....	103

4.2.3.2 $\Delta 2$: Intrasubject Intertrial VR simulator.....	105
4.2.3.3 $\Delta 3$, $\Delta 4$, and $\Delta 5$: Intersubject Intrasetting Comparisons.....	111
4.2.3.4 $\Delta 6$: Intrasubject Intersetting	117
4.2.3.5 Expert vs. Resident Comparisons.....	123
4.2.3.5.1 Interlevel Intrasetting OR.....	124
4.2.3.5.2 Interlevel Intrasetting Physical Simulator.....	127
4.2.3.5.3 Interlevel Intrasetting VR Simulator.....	130
4.3 Discussion.....	132
4.3.1 Context Comparisons.....	133
4.3.1.1 Intraprocedural Operating Room Variability.....	133
4.3.1.2 Intrasubject Intertrial VR Variability.....	134
4.3.1.3 Intersubject Intrasetting Comparisons.....	134
4.3.1.3.1 Operating Room.....	135
4.3.1.3.2 Virtual Reality Simulator.....	135
4.3.1.3.3 Physical Simulator.....	135
4.3.1.4 Intrasubject Intersetting Comparison.....	136
4.3.1.5 Interlevel Intrasetting.....	136
4.3.1.5.1 Operating Room.....	137
4.3.1.5.2 Virtual Reality Simulator.....	137
4.3.1.5.3 Physical Simulator	138
4.3.1.5.4 Experts vs. Residents.....	138
4.3.2 Performance Measure Reliability.....	139
4.4 Conclusions.....	139
Chapter 5: Conclusions and Recommendations.....	141
5.1 Introduction.....	141
5.2 Review of Research.....	141
5.2.1 Experimental Surgical Tool.....	141
5.2.2 Data Collection.....	143
5.2.2.1 The Operating Room.....	143
5.2.2.2 The Experimental Surgical Tool.....	143
5.2.2.3 Simulators.....	144
5.2.3 Data Fusion.....	144
5.2.4 Performance Measures.....	145
5.2.5 Context Comparisons.....	146
5.2.6 Simulator Validation.....	147
5.3 Recommendations.....	147
5.3.1 Software.....	148
5.3.2 Hardware.....	148
5.3.3 OR Data Collection.....	149
5.3.4 Simulators.....	149
5.3.5 Other Recommendations.....	149

5.4 Partner & Future Studies.....	149
List of Terms.....	153
Bibliography.....	155
Appendix A: OR Study Experimental Protocol and Data Acquisition Procedures.....	165
Appendix B: Operational Definitions.....	169
Appendix C: University of British Columbia CREB approval.....	173
Appendix D: Medicine Meets Virtual Reality Conference Submission.....	175
Appendix E: SAGES Conference Submission.....	177
Appendix F: Transfer of Training from Simulator to Operating Room.....	179

List of Tables

Table 1.1 – Types of validity definitions.....	13
Table 2.1 – Criteria for design of the sensor mounting bracket	29
Table 2.2 – Advantages of data fusion	44
Table 3.1 – Performance measures available from the three contexts.....	72
Table 4.1 – Summary of successful data collection from each context.....	95
Table B.1 – Hierarchical subtask dissection definition.....	171
Table B.2 – Kinematics and force performance measures.....	172

List of Figures

Figure 1.1 – Typical minimally invasive surgery operating room setup.....	3
Figure 1.2 – A typical laparoscopic cholecystectomy operation.....	4
Figure 1.3 – Reduced DOF of motion of the MIS tool tip.....	5
Figure 1.4 – Physical and VR simulators.....	8
Figure 1.5 – Performance Measures.....	15
Figure 1.6 – Are laparoscopic surgical simulators valid?.....	18
Figure 2.1 – Maryland dissector tip.....	21
Figure 2.2 – Tool tip reference frame.....	22
Figure 2.3 – NDI Polaris optoelectronic position tracking system.....	23
Figure 2.4 – MDMArray.....	24
Figure 2.5 – Polhemus Fastrak magnetic position tracking system.....	27
Figure 2.6 – F/T system of Rosen (1999).....	28
Figure 2.7 – ATI Mini 40 F/T transducer.....	28
Figure 2.8 – Two cut views of a typical laparoscopic tool shaft.....	29
Figure 2.9 – Sensor mounting bracket on surgical tool.....	30
Figure 2.10 – Force/Torque sensor bracket two segments	30
Figure 2.11 – Force load path through sensor bracket and F/T sensor.....	31
Figure 2.12 – Laparoscopic trocar.....	32
Figure 2.13a – Overall view of the tool, which is then split into 3 sections for FBD analysis.....	32
Figure 2.13b – Effective tip forces & moments.....	34
Figure 2.13c – Free body diagram of distal end of surgical tool and force sensor.....	34
Figure 2.13d – Free body diagram of force sensor and stationary tool handle.....	35
Figure 2.13e – Free body diagram of tool handle and strain gauges.....	36
Figure 2.14 – Strain gauge circuit diagram for half-bridge circuit.....	37
Figure 2.15 – Strain gauges mounted on tool handle.....	38
Figure 2.16 – Mechanics of laparoscopic tool.....	38
Figure 2.17 – Interaction between strain gauges and force sensor.....	39
Figure 2.18 – Results from one calibration test.....	40
Figure 2.19 – Friction in the surgical tool handle and bracket.....	41
Figure 2.20 – Grip compensation.....	42

Figure 2.21 – Data fusion steps.....	46
Figure 2.22a – Noisy magnetic data.....	47
Figure 2.22b – GCV smoothed magnetic data.....	48
Figure 2.23 – Laboratory data.....	49
Figure 2.24 – Interpolate the magnetic data.....	50
Figure 2.25 – Interpolated magnetic data.....	51
Figure 2.26 – Difference curve.....	52
Figure 2.27 – Interpolated difference curve.....	53
Figure 2.28 – Fused data.....	54
Figure 2.29 – Computer generated magnetic and optical data.....	56
Figure 2.30 – RMS error for computer-generated data.....	57
Figure 2.31 – Laboratory collected magnetic and optical data.....	58
Figure 2.32 – RMS error for laboratory collected data.....	59
Figure 2.33 – Real OR magnetic and optical data.....	60
Figure 2.34 – RMS error for OR data.....	61
Figure 3.1 – Diagram of goals for this project.....	67
Figure 3.2 – Tool tip reference frame.....	73
Figure 3.3 – Components of the performance measurement system.....	74
Figure 3.4 – Custom designed data acquisition software.....	76
Figure 3.5 – Data post-processing.....	78
Figure 3.6 – Data synchronization process.....	80
Figure 3.7 – Strain gauge data with electrocautery noise.....	83
Figure 3.8 – Raw and noise removed strain gauge data.....	84
Figure 3.9 – Electrocautery affected magnetic data.....	85
Figure 3.10 – Electrocautery affected F/T data.....	86
Figure 3.11 – VR simulator vs. Physical simulator vs. OR.....	87
Figure 3.12 – Kolmogorov-Smirnov CPD.....	89
Figure 3.13 – Moving Block Bootstrap.....	91
Figure 3.14 – Confidence intervals for D-values.....	92
Figure 3.15 – CPD of D-values.....	93
Figure 4.1 – Context comparisons for surgical residents.....	96
Figure 4.2 – Interlevel context comparisons for expert and residents.....	97

Figure 4.3 – Resident 1 intraprocedure OR CPD.....	100
Figure 4.4 – Resident 1 intraprocedure OR D-values.....	101
Figure 4.5 – Resident 2 intraprocedure OR CPD.....	102
Figure 4.6 – Resident 2 intraprocedure OR D-values.....	103
Figure 4.7 – Resident 3 intraprocedure OR CPD.....	104
Figure 4.8 – Resident 3 intraprocedure OR D-values.....	105
Figure 4.9 – Resident 1 intertrial VR simulator CPD.....	106
Figure 4.10 – Resident 1 intertrial VR simulator D-value comparisons.....	107
Figure 4.11 – Resident 2 intertrial VR simulator CPD.....	108
Figure 4.12 – Resident 2 intertrial VR simulator D-value comparisons.....	109
Figure 4.13 – Resident 3 intertrial VR simulator CPD.....	110
Figure 4.14 – Resident 3 intertrial VR simulator D-value comparisons.....	111
Figure 4.15 – Intersubject intrasetting (OR) CPD.....	112
Figure 4.16 – Intersubject intrasetting (OR) D-value comparisons.....	113
Figure 4.17 – Intersubject intrasetting (VR simulator) CPD.....	114
Figure 4.18 – Intersubject intrasetting (VR simulator) D-value comparisons.....	115
Figure 4.19 – Intersubject intrasetting (physical simulator) CPD.....	116
Figure 4.20 – Intersubject intrasetting (physical simulator) D-value comparisons.....	117
Figure 4.21 – Resident 1 intersetting CPD.....	118
Figure 4.22 – Resident 1 intersetting D-values.....	119
Figure 4.23 – Resident 2 intersetting CPD.....	120
Figure 4.24 – Resident 2 intersetting D-values.....	121
Figure 4.25 – Resident 2 intersetting CPD.....	122
Figure 4.26 – Resident 3 intersetting D-values.....	123
Figure 4.27 – Lumped interlevel OR CPD.....	124
Figure 4.28 – Interlevel OR individual CPD.....	125
Figure 4.29 – D-values for the two experts and three residents in the OR.....	126
Figure 4.30 – Lumped interlevel physical simulator CPD.....	127
Figure 4.31 – Interlevel physical simulator individual CPD.....	128
Figure 4.32 – D-values for the two experts and three residents in the physical simulator....	129
Figure 4.33 – Lumped interlevel VR simulator CPD.....	130
Figure 4.34 – Interlevel VR simulator individual CPD.....	131

Figure 4.35 – D-values for the two experts and three residents in the VR simulator.....132

Figure 5.1 – New performance measures.....145

Figure 5.2 – Concurrent research projects at the Neuromotor Control Laboratory.....151

Figure A.1 – University of British Columbia operating room experimental set-up.....167

Figure B.1 – Five levels of the hierarchical decomposition.....169

Figure B.2 – Five phases of laparoscopic surgery.....169

Figure B.3 – Stage level diagram for cystic duct dissection (CDD) and gallbladder dissection (GBD).....170

Acknowledgements

I would like to acknowledge and thank all those who have supported me from the first day I began this journey. Firstly, my thanks to my supervisor, Dr. Antony Hodgson, for his enthusiasm and optimism in this project, even when things were at their bleakest. I could always count on him to put a positive spin on every aspect, and for this I am most grateful. I am most appreciative to the participating surgical residents that I harassed endlessly for their time, when their time is so limited. Thanks to Dr. Ed Chang, Dr. Naisan Garraway, and Dr. Kathy Hsu. Thanks also to Dr. Hamish Hwang for twisting the arms of his resident friends to participate in this project. I would also like to acknowledge Dr. Alex Nagy and Dr. Neely Panton for sharing their knowledge and expertise, and supervising the surgical residents in the operating room. Many thanks go to Marlene Purvey and her surgical staff, and also to Betty Whincup and the staff at the Sterile Supply Department.

Next I want to give a big high-five to Catherine Kinnaird and Iman Brouwer. You guys are the best! No one else understands as well as you what we went through to finish this project. As a wise Iman once said, "I thought this day would never come." I also need to give a big pat on the back to the orthopod girls: Stacy Bullock, Carolyn Sparrey, Carolyn Greaves and Christina Niosi. Thank goodness for coffee break is all I have to say about that. Also, thanks to the rest of the NCL crew. Keep up the good work! I also want to say thanks to Val Roy for being such a good suburb buddy.

I am very grateful to my friends at CESEI: Ferooz, Vanessa, Marlene, Humberto and Dr. Qayumi. Thanks for the comic relief, the fabulous printer/copier, and all the food. The students that get to work with you next are very lucky indeed.

Many thanks go to Ryan Jennings for being at home to lend an ear to listen to my whining, for pushing me constantly to work harder, and for never letting me give up. To my Dad whom I got good advice from about completing graduate studies, and giving continued support and encouragement. And lastly, thanks to my Mom, whom I could not have done this project without. I am eternally grateful to for her unwavering support, no matter what.

Chapter 1

Introduction and Literature Review

1.1 Introduction and Objectives

Minimally invasive surgery is an increasingly popular procedure that uses smaller incisions, and results in much shorter recovery periods for the patients. Unfortunately, the surgery is substantially more demanding for the surgeon, who must learn a new set of skills; to use long instruments inside the body, while looking at a monitor outside of the body. Simulators offer the surgeon an opportunity for unlimited practice, and for practice on unusual cases. In order for the training to be useful, the simulator must accurately reflect the skill set required in surgery. The goal of this project was to validate both a physical and a virtual reality simulator in terms of the kinematics and the forces used in comparison to those used during surgery.

Surgeons must learn to operate both with skill and safety. The use of surgical simulators has become more widespread and important in the training of surgical residents. It is important that researchers direct their efforts into the areas that are of most significance to the patients and surgeons alike. Objective measurements of a surgeon's performance are more readily available in a simulator as compared to taking measurements during a live operation. This is also important when evaluating trained surgeons. Making these measurements in a simulated setting would be ideal as it is much more easy to evaluate performance in a simulator than in the OR. New tool designs and improvements could also be tested in a simulator saving operating room time and money.

Surgical education has lagged behind other educational areas where simulators are commonplace for teaching and training novices. Other professions, such as aviation, have successfully included simulation training into their educational programs. The success in the pilot training industry has pushed surgical educators to continue research in this area. In a survey in 1999, 92% of program directors agreed that there is a need for technical skills training outside of the OR (Haluck 2001). This is a definite explicit sign that it is imperative that other methods of surgical education be explored.

The overall objective of our lab research was to create and apply a quantitative method of surgical performance in order to assess two laparoscopic surgical simulators. The shorter-term goals included a study of the validity and reliability of these surgical simulators, and a study of the minimum technological requirements of a virtual reality surgical simulator. We aimed to establish whether these simulators are reliable measurement devices.

The primary objective of the work presented in this project was to assess the validity of both virtual reality and physical laparoscopic surgical simulators. The second goal was to develop a new experimental tool and system capable of the collection and analysis of the performance measures used in the simulator validity assessments. Operating room data was compared to analogous tasks in the simulator settings. The new methods provide a standard for future simulator assessments.

1.2 Minimally Invasive Surgery

Minimally invasive surgery (MIS) has become a routine and usual method of performing many types of surgical procedures. MIS is also known as minimal access surgery (MAS) or keyhole surgery. Because of advances in technology and medicine, many open surgical procedures can now be performed using MIS.

The notion of MIS first began in the early 20th century (Nagy, 1992). After World War II, the two most important inventions related to endoscopy and MIS were developed: the rod-lens system and fibreoptics. After much development in surgical technique and camera technology, the first laparoscopic cholecystectomy using video was performed on a human in 1987 in Lyons, France (Mishra 2004). Within that year, many other surgeons were practicing their first cholecystectomies on humans on both sides of the Atlantic. Since the late 1980s, MIS has become commonplace in modern general surgery. The use of MIS in the United States in abdominal surgical procedures has reached 60-80% (Taylor 1995). A typical minimally invasive surgery operating room set-up can be seen below in Figure 1.1.



Figure 1.1: *Typical minimally invasive surgery operating room set-up. Notice the video monitors in the background and situated around the OR. The surgeons rely on these monitors to view the surgical field within the patient. The monitors show a direct video feed from the laparoscopic camera.*

Laparoscopic surgery has allowed surgeons to perform many of the same procedures as in traditional open surgery, but using small incisions (5-15 mm) instead of large abdominal incisions (7-15 cm) (Huntsville 2002). This increased use of MIS techniques throughout the years has led to benefits for patients. Studies have shown that the patient benefits in terms of reduced post operative pain, smaller scars, reduced hospital stay, quicker return to normal physical activities and therefore, a quicker return to work (Treat 1996, Périssat 1995). It is common, and proven to be safe, for routine cholecystectomy procedures to be day surgeries, with the patient coming into the hospital in the morning and leaving for home in the afternoon (Prasad 1996). Other patients are discharged from the hospital usually 1 or 2 days after the cholecystectomy with low complication rates (Lujan 1998). A typical laparoscopic cholecystectomy operation set-up can be seen in Figure 1.2.

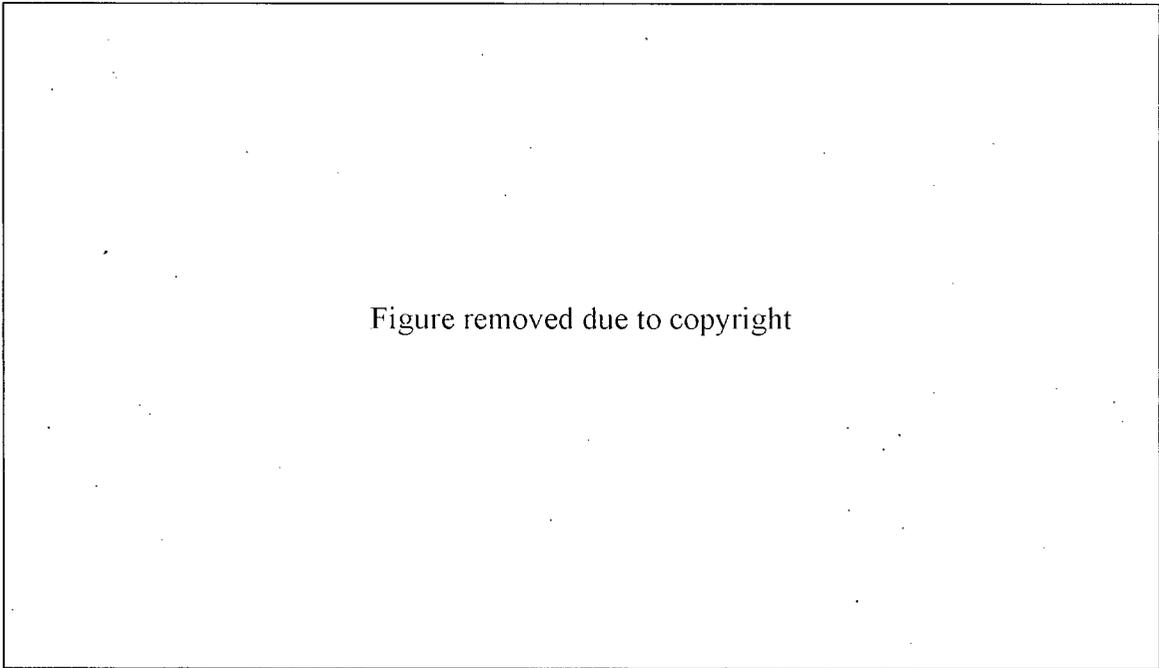


Figure removed due to copyright

Figure 1.2: *A typical laparoscopic cholecystectomy operation.*

Although there are many obvious benefits for the patients in MIS, there is a specialized skill set required by the surgeon that is much different than in traditional open surgical techniques. Laparoscopic tools very often limit the surgeons' dexterity and range of motion, and surgeons use uncomfortable postures to complete tasks (Person 2001). Laparoscopic tools are considered ergonomically poorly designed, awkward and not easy to use (Berguer 1999, Emam 2001, Treat 1996). Also, the time to complete a laparoscopic procedure compared to the same open surgical procedure can be up to 30% longer (Glinatsis 1992, Treat 1996). Conversely, other studies have shown that there is either no significant difference in surgical times, or that the laparoscopic approach may actually be shorter in time duration (Pessaux 2001).

1.2.1 The Challenges of MIS for Surgeons

The special skill set required of surgeons for laparoscopic surgery is especially difficult for the trainee to learn. One of the aspects that a novice surgeon must adapt to is what is known as the fulcrum effect (Jordan 2001). Specifically in laparoscopic surgery, this is when the surgical tool is inserted into the abdomen, creating a fulcrum. The surgeon experiences a motion

reversal. For example, when the surgeon moves their hand to the left on the outside of the body, the tool tip is moving to the right inside the abdomen. This is a basic motor skill that novice surgeons must learn.

Another issue is video-hand-eye coordination (Ballantyne 2002, Perkins 2002). The surgeon is no longer directly viewing the surgical field, but rather a 2D video monitor of what is happening inside the abdomen. The surgeon is working with their hands outside of the abdomen using longer surgical tools than in open surgery, watching a video feed of what the surgical tools are doing inside the abdomen. This lends itself to a lack of depth perception and makes tasks such as suturing and knot tying more difficult. Tactile feedback is also reduced in MIS creating yet another problem for surgeons to overcome.

In laparoscopic surgery, there are a reduced number of degrees of freedom (DOF) for the surgical tool. The laparoscopic tool only has 4 DOF as opposed to the open surgical tool, which has 6 DOF. This limits the surgeon's dexterity and range of motion (Tendick 1995, Ballantyne 2002). The tip movement is limited to pitch, yaw, roll and plunge (i.e. in/out of the abdomen), as shown in Figure 1.3 (Person 2000).

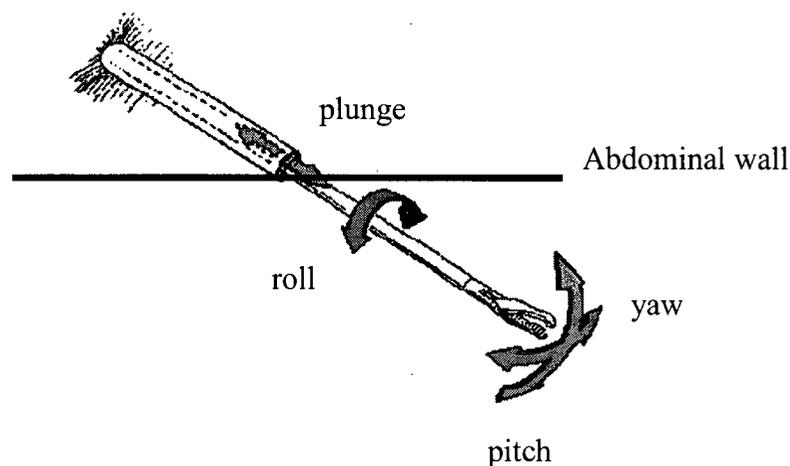


Figure 1.3: *Reduced DOF of motion of the MIS tool tip. DOF are roll, pitch, and yaw about the fulcrum created by the entry portal and plunge through the portal. (Modified from source: Person 2000).*

Researchers are studying methods to deal with these limitations of laparoscopic surgery by looking into new technologies such as 3D vision systems (Jones 1996, McDougall 1996,

Chan 1997, Hanna 1998), robotic surgery (Dakin 2003, Hubens 2003, Ruurda 2003, Ruurda 2002, Vuilleumier 2003) telerobotic surgery (Ballantyne 2002, Marescaux 2001, Perez 2003), and interactive image guidance (Harms 2001, Herline 2000, Stefansic 2002).

1.2.2 The Challenges of MIS for Surgical Educators

Due to the inherent limitations of performing MIS, the surgical education community must face the challenge of deciding where to train the surgeons and how to evaluate them. These issues are of importance to the surgical community and public alike, in that it is imperative that surgeon trainees finish their education with the ability to operate safely and effectively.

Researchers unanimously agree that the current training and evaluation of surgeons is subjective, unreliable and costly (Feldman 2004, Lentz 2002, Rosser 1998, Winckel 1994).

This is one reason why there has been pressure to investigate the feasibility of using surgical simulators for the purposes of training and evaluation.

1.2.3 Reasons for Using Surgical Simulators

Surgical simulators have many possible useful applications. Surgeon certification and tool evaluation are just two of the possible uses of validated simulators.

1.2.3.1 Surgeon Certification

The ability to quantitatively assess surgical performance is important to the training and certification of both novice and expert surgeons. The methods that we have developed will allow performance measurement in the OR followed by a comparison to the surgical reference database of performance measures from surgeons of varying skill levels. This will allow for a quantitative analysis of skill level and a method for identifying where improvements are needed. For example, when a novice surgeon seems to be having difficulty in a certain task, this could ideally be identified, and advice given specifically to address the problem.

1.2.3.2 Equipment Design and Evaluation

Tool and equipment designers could evaluate the performance of their new instrumentation in a validated simulated environment. The designers could ideally be confident that the new tools will give the same performance in the OR. The evaluation of new tools would be an iterative

process where the new tool is compared to a reference database of performance measures for past tool designs.

1.2.3.3 Transfer of Training

If a simulator is shown to be valid (see Section 1.5 for further details), the next step in furthering the push for simulators to be used in surgical education programs is to determine the transfer of training issue. Do novice surgeons who practice in simulators show a significant improvement in the operating room? In other words, if a novice surgeon spends X amount of time practicing on a simulator, will there be a quantifiable improvement in OR performance, as opposed to a similar novice who does not have any simulator training?

The original goal of this project was to study the issue of transfer of training from simulator to human operating room, as this is a subject that needs analysis in the surgical education and simulator fields of study. Unfortunately due to many logistical nightmares such as patient recruitment, scheduling, and many others, this project was converted to a simulator validity study. This was the most logical step as the proper OR and simulator data had already been collected. Further information on the transfer of training from simulator to OR issue can be found in Appendix F.

1.3 Current Training Methods

Success in laparoscopic surgery is very dependent on the surgeon's proficiency and experience (Perissat 1995). The apprenticeship-training model is still the most commonly used for providing experience to surgical residents. This is basically where the surgical resident shadows the expert surgeon, and learns the tools and tricks of the trade by observation, questions, and some hands-on practice. The disadvantages of this approach are that the surgical educator has no control over which patients require surgery, potentially limiting a novice's contact to a small variety of cases. Consequently, the novice surgeon may only be exposed to a limited pool of anatomy and pathology.

The use of human cadavers as a training model has been used in surgeon training programs with some success (Martin 2003). However these cadaveric models have their own

disadvantages: they are expensive, subject to availability, have different tissue properties than a live human, and there is some concern over transmission of disease (Nelson 1990).

Animal models may avoid some of the stresses and time constraints of apprenticeship training, but the anatomy often differs from that of humans. The disease state cannot often be reproduced in the animal, and an animal care facility is expensive. There are many moral and ethical issues related to training on live animals. The United Kingdom has banned the use of animals for surgical training (Lirici 1997, Moorthy 2003).

Surgical simulators, both physical and virtual reality (Figure 1.4), are becoming more widely used and accepted for use in surgical education, although their use is still limited at the University of British Columbia surgical training program, in that currently there is no prescribed simulator training. The use of virtual reality (VR) systems with haptic (force-feedback) interfaces has garnered much interest. Simulators are designed to highlight either or both the psychomotor skills (e.g., clipping and suturing skills) and the cognitive aspects of surgery (e.g., decisions about the steps to follow during a procedure). Simulator training is safe, highly available and unlimited practice is possible. No supervision is necessary when a novice is using these simulations.

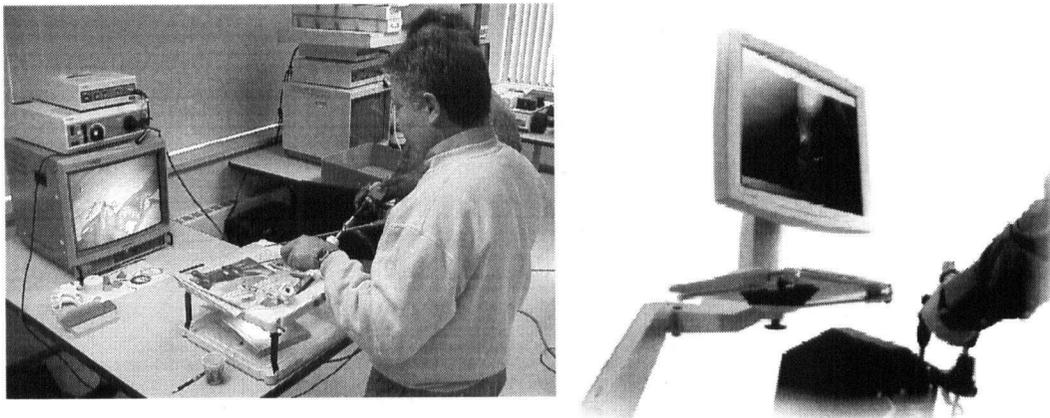


Figure 1.4: *Physical and VR simulators. The left picture shows a physical simulator using regular laparoscopic tools, and an inanimate model. The right picture shows a VR simulator with computer-generated models.*

Recently in the surgical education community, there has been quite some interest expressed specifically in virtual reality (VR) simulators. Most of the current studies were done with VR simulators. Although VR simulators are comparatively expensive in initial cost compared to

bench-top simulators, they do have advantages. These VR simulators can be programmed to include variant anatomy (pathologies, rare occurrences), temporal changes (patient status, bleeding), and of course, provide objective measurements (e.g., time, errors, kinematics, etc.) through the computer software.

Surgical educators have come to the realization that the surgical training programs should become more structured, and that surgical models and simulators should have a more important role in training, evaluating, and certifying surgeons (Feldman 2004). For the attending expert surgeons to be willing to change to this new paradigm of teaching, it is imperative to eventually demonstrate that time spent in a simulator can replace time spent in the operating room. This is not only important to the educators but to the hospital administrators and the taxpayers alike. In the US in 1997, the estimated cost of training 1014 general surgery residents in the OR was \$53 million (Bridges 1999). This cost was mostly attributed to the extra amount of time (2480 hours) spent in the ORs when a resident is operating. So it is quite obvious that financially, simulators may save time in the OR and therefore money in training surgeons.

1.4 Current Methods of Surgical Performance Assessment

A clear and objective method to assess performance and skill in laparoscopic procedures is potentially useful for many aspects of surgery including surgical resident evaluation, simulator validation, and surgical tool evaluation. Since the early 1970s, when Kopta developed one of the first methods for performance evaluation, the surgical education community has become quite interested in this topic (Kopta 1971). Current evaluation methods are known to be subjective and possibly unreliable, so there is a need for objective methods to measure surgical performance (Rosser 1998, Winckel 1994, Lentz 2002, Chung 1998, Feldman 2004).

One of the more commonly used methods for surgeon evaluation is the structured skills assessment form. These forms can be a type of checklist or a form where the evaluator must describe/fill-in specific areas. This type of form allows for a complete intra-operative performance evaluation, which can analyze both psychomotor and cognitive skills of a surgeon. Many researchers have used this type of evaluation in various studies (Winckel 1994, Eubanks 1999, Reznick 1997). Many studies have also been done to show the validity and

reliability of using these structured skills forms (Martin 1997, Goff 2001, MacRae 2000, Cohen 1990, Regehr 1998, Faulkner 1996). The shortcomings of these types of forms include patient variability, stress associated with the OR environment, and the difficulty of recognizing the level of technical skill. These surgical skills are not specifically quantified during these structured skills assessments.

Another very common measure used to quantify surgeon performance is the speed to complete a task. The time required to perform a procedure is easy to measure and has been used in many studies (Derossis 1998, Fried 1999, Hanna 1998, Hodgson 1999, Rosser 1997, Starkes 1998, Szalay 2000, Taffinder 1999).

Quality of performance has also been used as a method of evaluation. This measure is generally evaluated using subjective methods such as checklists and global assessments ratings (Eubanks 1999, Feldman 2004). Global assessment ratings are a type of subjective evaluation method where an evaluator can rate the subject on a scale (i.e. 1-poor to 5-excellent). Objective Structured Assessment of Technical Skill (OSATS) is one of the more commonly used and researched qualitative assessment techniques (Martin 1997). The OSATS is a set of operation-specific checklists that is specific to a physical simulator. Quality is a subjective performance measure that can usually be easily implemented into any type of evaluation method.

A measure of error has also been studied with some interest. Although most surgeons do not like to speak about errors or injuries occurring during surgery, errors and injuries do occur (Francoeur 2003, Way 2003). The methods of evaluating error vary from objective measures, usually in simulated settings, (Francis 2002, Grantcharov 2003, O'Toole 1999) to subjective observed measures (Bann 2003, Joice 1998, Seymour 2002).

Force/torques are another measure that can be analyzed. More recently, Rosen and colleagues successfully completed a study in a porcine model analyzing force/torque signatures on the surgical tool tip (Rosen 2001). The researchers used a Markov modeling method (method to detect patterns) along with a structured process to classify tool movements to evaluate surgical performance. They showed they could correctly categorize surgeons into two different experience levels (novice and expert) because of similarities derived from their Markov

models. Other researchers have also incorporated force measurements into their measurement and training systems (deVisser 2002, Hanna 1997, Morimoto 1997, O'Toole 1999, Wagner 2002, Verner 2002, Yamauchi 2002).

1.5 Simulator Validation

Validity is a general term with many definitions. The American Psychological Association developed a set of standard definitions to aid in validity studies (APA 1974). From these standards, we are most interested in behavioural correspondence validity (now referred to as validity), as this is how the human operator treats the simulator as compared to the real situation. By comparing the simulator and the real situation during analogous tasks in terms of human operator behaviour, this can be tested (Blaauw 1982).

It is of utmost importance that the simulators used for surgical skill training, assessment, and certification be validated. The test in validating surgical simulators is to prove that performance in the simulator will represent performance in the OR. To ensure a valid simulation, we must make certain that a surgeon treats the simulation, in as many applicable and quantifiable aspects as possible, the same way they treat a live patient.

Many research groups have put considerable time into validating the currently available surgical simulation systems (Adrales 2003, Bloom 2003, Feldman 2004, Paisley 2001, Schijven 2003, Strom 2003, Taffinder 1998). There are five common different levels of validity from least to most rigorous: face, content, construct, concurrent and predictive (Table 1.1).

Face validity is a type of validity that is assessed by experts' review of the contents of the simulator. It is a subjective test as it is based on expert opinion, and is usually done in the initial phases of validity testing. Content validity is an extension of face validity, where the expert would use a checklist to reduce the rater subjectivity. The content validity tests to see if the simulator contains the steps and skills that are used in the real procedure. These simple validity tests are also the most subjective.

Construct validity is tested by discrimination between skill levels. It tests the degree to which the simulator “identifies the quality, ability or trait it was designed to measure” (APA 1974). This is another common test applied to surgical simulators.

Concurrent validity is a validity test that correlates performance with the current gold standard. For surgical simulators, the gold standard is operating room performance by expert surgeons. Currently, the gold standard measurement is done with performance-specific checklists in the OR (Feldman 2004). Using this approach is generally time consuming and is still considered subjective.

Predictive validity is whether the simulator can predict actual performance in the real setting. This type of validity is rather controversial as decisions about junior surgeons may be based on simulator performance. If predictive validity is shown, a poor simulator performance may remove juniors from continuing in their surgical training (Gallagher 2003).

In a parallel project to the one to be described a fellow lab member, Catherine Kinnaird, investigated some aspects of validity of both physical and virtual reality surgical simulators with expert surgeon subjects (Kinnaird 2004). In Kinnaird’s work, a new type of validity, performance validity, was introduced. Performance validity is a quantitative assessment of measurable quantities of performance in the OR (i.e. kinematics and force profiles); if these measures are the same as in the surgical simulator, then the simulator can be considered valid. This new type of validity allows for objective assessments using the same measurable quantities in many different environments. Therefore, we have uniformity and consistency when making evaluations in the OR or simulators.

Table 1.1: *Types of validity definitions (Gallagher 2003)*

Validity	Definition	Studies
Face	Expert Opinion	Haluck 2001, McCarthy 1999
Content	Checklist of matching elements	Paisley 2001, Schijven 2002
Construct	Differentiates between skill levels	Adrales 2003, Datta 2002 Gallagher 2004, Grantcharov 2002, Taffinder 1998, Schijven 2003
Concurrent	Correlates with gold standard	Ahlberg 2002, Feldman 2004, Grantcharov 2004
Performance	Quantifiable performance measures same as “real” setting	Present study, Kinnaird 2004
Predictive	Predicts future results	N/A

1.5.1 Construct, Performance, and Concurrent Validity

A very important step in the evaluation of surgical simulators is to establish construct validity. Construct validity is a quality established when performance scores on a simulator reflect the ability of the person performing the actual procedure; therefore an expert should score higher than a novice. Different researchers have studied construct validity of various different types of simulators such as arthroscopy and gastrointestinal endoscopy (Bloom 2003, Srivastava 2004). The concept of construct validity is often regarded as an important central theme in validation studies (Gabberson 1997).

In the laparoscopic simulator field, there has also been extensive research into the validity of the MIST-VR simulation system (Mentice Medical Simulation AB, Gothenburg, Sweden). The construct validity of this particular system has been established in a few different studies (Gallaher 2002, Gallagher 2001, McNatt 2001). The latest study on the MIST-VR showed that the system has “discriminative validity” and was capable of evaluating the psychomotor skills necessary in laparoscopic surgery and discriminating experts and novices (Gallagher 2004). The MIST-VR system has been shown to discriminate between the performances of subjects with similar experience and similar skill levels. Subjects can then be grouped according to psychomotor skill level. Discriminative validity is a further refinement of construct validity.

Construct validity has also been shown in physical simulators such as the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS) system (Fried 2004). This was an in-depth study with over 200 participating surgeons and trainees in 5 countries. The MISTELS system is the physical simulator used by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) Fundamentals of Laparoscopic Surgery (FLS) program.

The current “gold standard” for concurrent validity studies is OR performance. But the problem with this is the subjective methods (i.e., checklists) to evaluate this OR behaviour.

1.6 Research Question

Because the previous methods for investigating validity in surgical simulators have been done with subjective assessments, there is a need to further the study into simulator validity by using quantitative measures. What we would like to know is whether or not motor behaviour in the simulator is analogous to the OR. This will allow us to determine whether the simulator is a good training and evaluation environment.

In a complementary study to this project, Catherine Kinnaird (2004) began the investigation into simulator validity by evaluating expert surgeons in the OR and with both physical and VR simulators. That study looked at the performance validity of these simulators by comparing data from the OR with that of the simulators. This expert surgeon study led us to want to investigate further the validity of these simulators.

Therefore, the project to be described in this manuscript is a furthering of the validity study of these simulators. The primary objective of this project was to investigate the performance, construct, and concurrent validity of both a physical and VR surgical simulator. The construct validity study used the expert surgeon data analyzed by Kinnaird (2004). The secondary objective was to develop a system that was capable of collecting and analyzing quantitative data from the human OR.

1.7 Developing a Quantitative Assessment Method

The development of a quantitative method to assess surgeons in the human OR required much thought and preparation. To be able to study the validity of surgical simulators and gather the performance measures that will allow for various context comparisons, we needed to improve and elaborate upon performance measures previously established within our lab. The performance measures that have been used previously include time, kinematics, joint angle and event sequencing (McBeth 2002). As shown below in Figure 1.5, known in our lab as the “Wheel of Performance”, there are other measures that can be made, and incorporated into our system of performance evaluation.

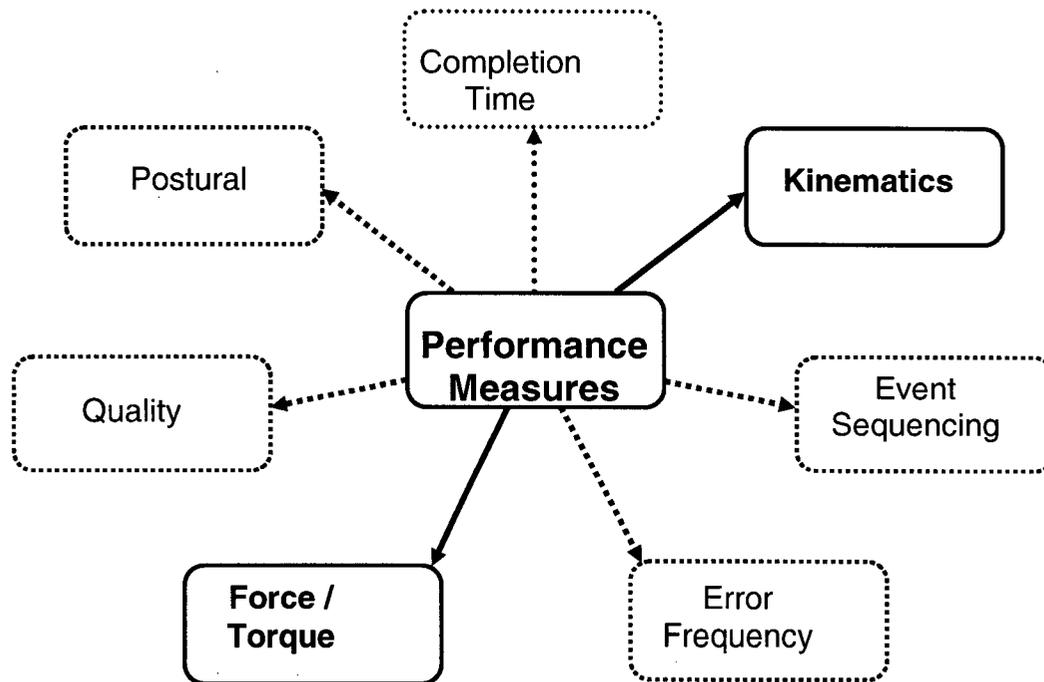


Figure 1.5: *Performance Measures.* The performance measures in bold and in the solid-line box are the ones used in this study. The measures shown in the dotted boxes will not be specifically studied in this thesis.

Due to the time constraints of this project, we focused our study on the following measures: kinematics, and force/torque. The performance measure of quality can be included easily by including checklists/questionnaires of some type, and as mentioned above, postural and event sequencing has been successfully completed in a previous study in our lab.

1.7.1 Kinematics

For this study, we have continued the work by McBeth (2002) to gather and analyze kinematics data for the surgical tool tip during laparoscopic surgery. We required a high frequency tracking system that would give us three-dimensional position and orientation data. For this type of tracking, there are many types of commercial systems available such as optoelectronic, magnetic, ultrasonic, and each system has its own advantages and disadvantages.

Optoelectronic systems can provide wireless high frequency data, and can be sterilized for OR use. The hybrid systems that are able to track both passive (wireless) and active (infrared) markers are useful in many circumstances. Disadvantages include line-of-sight problems and interference with external infrared sources. McBeth (2002) used an optoelectronic system and did have problems with line-of-sight where some procedures had virtually no usable data, which led to unreliable results. The optical system also had a low sampling frequency (30Hz), and this led to difficulties in producing velocity, acceleration and jerk profiles. In this project, we wanted to improve upon McBeth's method, and produce high frequency continuous kinematics data.

A tracking system that was available to us that seemed to overcome the problems of the optoelectronic system was an electromagnetic system. Electromagnetic sensors give higher frequency data sampling and are not affected by line-of-sight issues, but a wire to the interface unit connects each receiver. External ferrous materials and electromagnetic fields also detrimentally influence these sensors. Because of these issues, an electromagnetic system could not be used solely in the OR environment.

The study created a kinematics data collection system that incorporates both the optoelectronic and electromagnetic tracking systems. This overcomes the line-of-sight and low sampling rate problems of the optical sensor, and the low accuracy, metal interference problems of the magnetic sensor. By using a combination of the two position sensors, we are able to achieve a continuous high frequency kinematics dataset.

1.7.2 Force/Torques

To measure forces and torques, again there are commercially available systems. Force and torque measurements are made with specially designed force/torque sensors. For use in this project, it was important that we find a sensor that was small enough, yet robust enough, to be used in the OR. Strain gauge based force sensors are commonly used, and easily available, and can be gas sterilized for use in the OR. We followed the lead of Rosen (2001) with their technique of mounting a force sensor onto the shaft of a surgical tool. We also used strain gauges mounted to the surgical tool handle to aid in the calibration of the force sensor and to measure grip forces.

1.8 Project Goals

Surgery and surgical education are at a point where the traditional “see one, do one, teach one” teaching technique is no longer acceptable. Surgical education experts have more recently looked into the possibility of using simulators to train, test and certify surgeons. Before these simulators can be used in widespread practice, a thorough evaluation of the systems must be done. Validation of these physical and virtual reality simulators is of utmost importance, as a valid simulator will provide an environment that closely approximates the environment where the task will eventually be performed (Prystowski 1999).

The primary goal of this project is to assess construct and performance validity of two surgical simulators: virtual reality and physical (Figure 1.6). Construct validity refers to the concept that the context actually recreates the environment that it intends to recreate. A method of testing this in a surgical setting is to see whether expert surgeons perform better in these simulators than resident surgeons. A simulator that shows construct validity will be able to detect the skill level differences between experts and novices. Performance validity of a simulator is where the simulator’s behaviour is the same as in the OR. If a subject performs the same quantitative measures (such as kinematics or force) in a simulator as in the OR, the simulator is said to show performance validity. In turn, we also begin an investigation for quantitatively assessing concurrent validity of the both VR and physical simulators. We are able to make a quantitative “gold standard” measurement in the OR with expert surgeons (data analyzed by Kinnaird 2004), and gather the same performance measures in all other contexts (i.e. both VR and

physical simulators). The results from this study could then be used in the design of new simulators, surgical tools and techniques, surgeon training and evaluation.

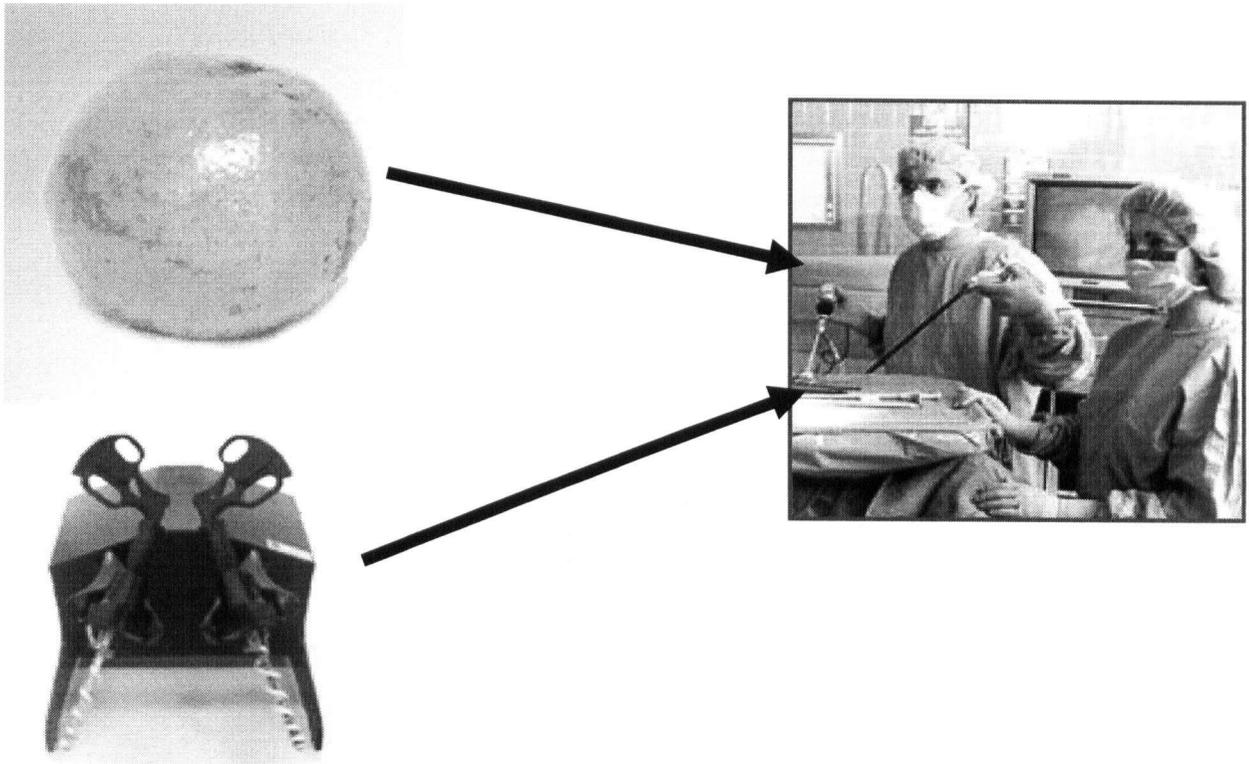


Figure 1.6: *Are laparoscopic surgical simulators valid? We are looking for similar motor behaviours between the simulator and the OR to investigate validity of laparoscopic surgical simulators. The orange represents a physical simulator, where the task was to peel the skin off the orange and remove a few segments; the bottom left picture is of a VR simulator interface.*

As mentioned previously, our performance measures of kinematics and forces were used for our quantitative measures for our validity study. We required a continuous high frequency signal in both these measures, and our existing lab system did not allow for this (i.e., occlusions in optical data). Therefore, the secondary goal of this project was to develop a new tool that would allow for us to get these continuous high frequency measures. The new data collection and analysis system incorporates a data fusion of the two kinematics data streams that eliminates the problem of occluded optical data. Previous methods of combining kinematics data done in the surgical environment have been attempted (Birkfellner 1998, Nakamoto 2000) but none of them were a true fusion of kinematics data.

Chapter 2 describes the design of the new experimental surgical tool, and the design of all the subsystems required for data collection and analysis. It provides a thorough description of the data fusion technique of two kinematics data streams to create high frequency continuous performance measures from the gathered data in the OR and physical simulator as well as the force measurement considerations and calibrations that are required for extraction of force performance measures.

Chapter 3 is the experimental methods used to collect data in the OR, and with the two simulators. A description of the equipment used and details of the data post-processing are also included.

Chapter 4 contains the results of the experimental testing and a discussion of these results. The reliability of the chosen performance measures and the subject and context variability relating to validity of the surgical simulators is investigated.

Chapter 5 is a summary of the findings, conclusions and recommendations for future work. The conclusions relate to current and complementary studies that affect studies in surgical education and simulation.

Chapter 2

Experimental Laparoscopic Surgical Tool for Performance Measure

Assessment

2.1 Introduction

Minimally invasive surgery is now a common and essential component of modern surgical medicine. Unfortunately, the same developments in surgical education and assessment have not kept pace. The current methods of surgical assessment have been shown to be subjective and unreliable (Chung 1998, Feldman 2004, Lentz 2002, Rosser 1998, Winckel 1994). Therefore, it is agreed there is a need for an objective method to assess surgical performance.

For many years, the notion that operative skills should be evaluated has been brought up repeatedly (Kopta 1971). Surgical simulators may provide an excellent venue for performance evaluation, as the measures can be objectively measured. Bench-top trainers, virtual reality (VR) systems and animal models are all used in surgical education programs currently. The performance measures that have been used by researchers include completion time, errors, force/torque signatures, event sequencing, and tool tip kinematics (Chung 1998, deVisser 2002, Derossis 1998, Hanna 1998, McBeth 2002, Rosen 2002, Way 2003, Yamuchi 2002).

The longer-term goals of the lab projects are to create a surgical skills database where surgeons could look-up their performance as compared to others. A surgical resident would be able to compare their performance to others of their own level, and see what needs to improve, or where they excel. But in order to do this, research must be done to validate the surgical simulators, and prove that training in a simulator does improve OR performance. Currently, studies have shown that expert surgeons perform better in simulators than novices, and practicing in a simulator leads to improvement in the simulator (Derossis 1998, Fried 2004, Rosser 1997). It has also been shown that assessments made in a simulator can be used to monitor progress (Derossis 1999, Fried 2004), and that practice in a porcine model leads to OR performance improvement (Fried 1999). And even more recently, breakthrough studies have shown that practice in a simulator would indeed lead to improvements in the human OR (Seymour 2002, Grantcharov 2004). This tells us that skills learned in a simulator could be used to replace OR time for learning.

This chapter describes the design and considerations for creating a new tool and data collection system to measure OR and simulator data used in studying the validity of both physical and virtual reality simulators. The objective is to improve upon the current tools used to gather the performance measures, and to add measurement of force to the system originally created by former lab member Paul McBeth (2002). A new technique was also created to fuse our two gathered streams of kinematics data to create a high frequency and continuous kinematics data stream.

2.2 Laparoscopic Surgical Tool

The laparoscopic surgical tool that was used in these studies was a Maryland dissector as seen in Figure 2.1. This particular tool was chosen as it is used the most during the initial parts of the laparoscopic cholecystectomy procedure to dissect away the surrounding tissues from the cystic duct and artery. It is used to pull, spread, and tear away the extraneous body tissues. When it is connected to the electro-surgical unit, it is capable of burning and cauterizing tissues. This particular tool was chosen on recommendation of an expert surgeon participating in our studies.

We obtained a commercially available tool through Storz Endoscopy. These tools have an interchangeable tool tip insert. Other tool tips may be purchased and used instead of the Maryland dissector insert. This is a good feature for future work, as different tips and therefore motions and forces will be available for data collection.

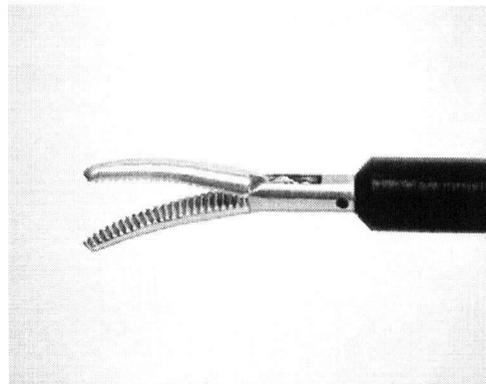


Figure 2.1: *Maryland dissector tip. Used for dissecting away surrounding tissues.*

2.3 Performance Measures

There are a wide range of performance measures that are available for assessing surgical skill. In consultation with expert surgeons, literature searches, and following the protocol from the previous study done in our lab by Paul McBeth (2002), we are continuing with the chosen performance metric of tool tip kinematics, and with the addition of tool tip force/torque. We are no longer including completion time, ergonomics/joint angles and event sequencing that were previously completed, but they can easily be re-implemented back into the system. The following sections describe further the selected performance metrics and the methods we used to collect this data.

2.3.1 Kinematics

The use of tool tip kinematics measures in assessing surgical performance has become more common in surgical performance measurement systems (McBeth 2002, Rosen 1999). Rosen's group has created the BlueDragon system, which measures kinematics of the tool tip *in vivo* in a porcine model (Rosen 2002). Another group has incorporated electromagnetic trackers to measure distance, number of movements, and speed for a surgeon's hand movements in a laboratory setting (Taffinder 1998, Smith 2002). In a previous study within our lab by McBeth (2002), kinematics data was collected using an optoelectronic position tracking system. Our group continued with McBeth's work and further elaborated and improved the system to measure tool tip kinematics data to investigate tool tip velocities, acceleration, and jerk in the following tool tip directions: axial, grasp, translation, transverse, absolute and roll about the tool axis (Figure 2.2).

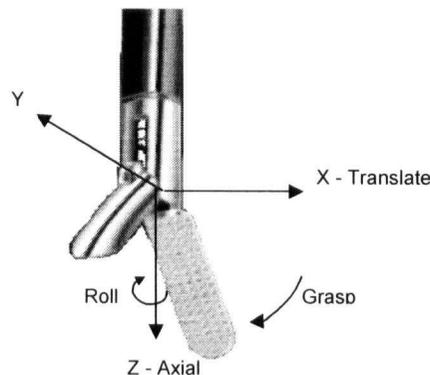


Figure 2.2: Tool tip reference frame. Tool tip directions with respect to the tool handle. The axial (z) direction is along the tool shaft. The grasp (y) is in line with the tool jaws. The translate (x) direction is in the perpendicular direction of the y and z axis.

2.3.1.1 Optoelectronic Position Tracking

In a previous study in our lab by Paul McBeth (2002), an optoelectronic motion tracking system was used to collect the kinematics data. According to the product manual, the Northern Digital (NDI Northern Digital Inc., Waterloo, ON, Canada) Polaris Hybrid Tracking System (Figure 2.3) is capable of tracking the 3D positions of both infrared light emitting diodes (IRED's) and passive reflective markers with an accuracy of $\sim 0.2\text{-}0.3\text{mm}$. In our study, we only used passive markers. This optoelectronic system was originally chosen because surgeons have seen them in the operating rooms and are familiar with their presence, the parts are easily sterilizable, a system was available, and we were primarily interested in postural data.

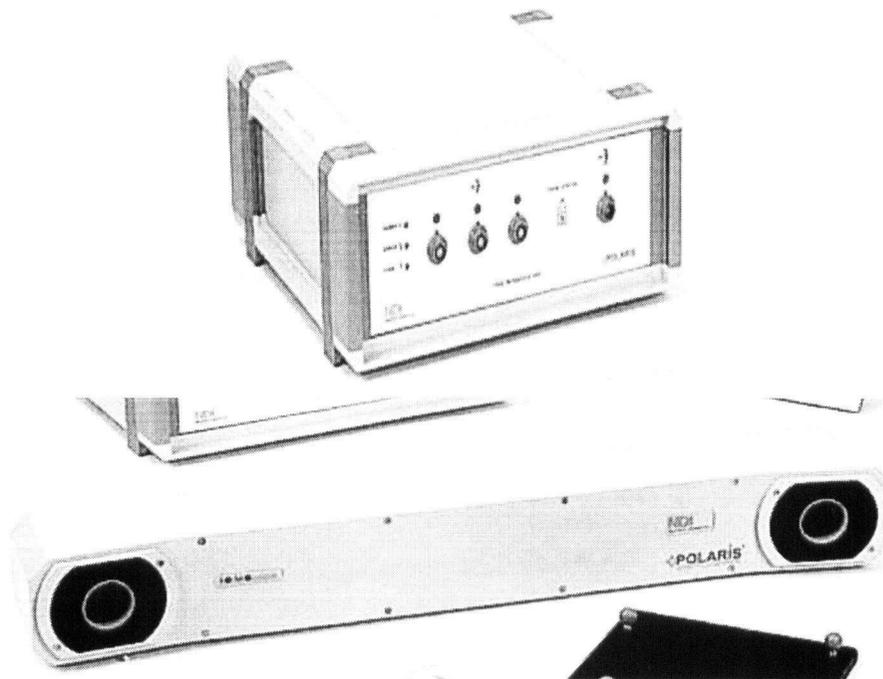


Figure 2.3: *NDI Polairis optoelectronic position tracking system. The top picture is the camera unit, and the lower picture is the tool interface unit.*

The Polaris system uses an infrared camera to track the desired markers. It requires three passive markers (retro-reflective balls) to establish an array, or reference frame. Polaris records the position and orientation of the reference frame with respect to the camera. A Multi-Directional Marker Array (MDMArray) was custom designed and made by McBeth (2002), and was attached to the experimental tool to track tool movement. This specially designed array was created to make the tool visible from many angles compared to just a standard planar array. The standard array was one of the original problems with this system.

The MDMArray has five geometrically unique faces, and can be rotated in many directions to allow the Polaris camera to track one face at a time. This allows for improved visibility of the passive markers to the camera and more continuous data to be collected, as intermittent data and therefore gaps in the data, is a significant problem. See Figure 2.4 for a picture of the MDMArray.



Figure 2.4: *MDMArray. Halo of optical passive marker balls used for optical position tracking. The infrared camera tracks faces (3 balls) of the array.*

The study conducted in our lab previously has shown that the Polaris optoelectronic system is usable in the OR, but some limitations were discovered. Because the Polaris depends on line-of-sight from the camera to the marker arrays, these arrays can become occluded from the camera's view by surgeon movements, interrupting and leaving gaps in the data stream. It was found that during typical manipulation tasks, the clipping tool was visible 78 +/-12% of the time even with the MDMArray (McBeth 2002).

2.3.1.1.1 Other Kinematics Options

Because our goal was to have a system that could gather continuous high frequency data to obtain our performance measures, we considered various options such as:

- Re-designing/modifying the current marker array to allow for more positions of the array to be seen
- More optical tracking cameras to allow the arrays to be seen from multiple angles, therefore increasing visibility
- Incorporating of a second motion tracking system:
 - Accelerometer/gyro
 - ShapeTape™ (flexible tape like position sensor which reports its shape)
 - Electromagnetic system

The options of changing the marker array or adding more cameras to the optical tracking system were discarded as this may improve the problem of occlusions/gaps, but would likely not solve it completely. Also the sampling frequency would still remain relatively low. The accelerometer/gyro option was considered, but was not easily available in our lab, and the same for the ShapeTape™. Neither of these systems would have been reasonable to design and debug in a reasonable amount of time. The electromagnetic system was available for our use as it was available through inter-departmental collaborations, and would provide the high frequency and continuous data stream that we required.

2.3.1.2 Electromagnetic Position Tracking

Electromagnetic tracking systems have been used in the past in surgical applications, but problems such as electrical noise and interference have been reported (Datta 2002, Frantz 2003, Smith 2002).

Other researchers have attempted to combine sensors in the surgical environment. A study completed by Birkfellner and colleagues (Birkfellner 1998) at the University of Vienna successfully combined and calibrated a hybrid (optical and electromagnetic) tracking system. Their motivation for merging the two tracking systems was similar to ours in that they were concerned with the optical system's line-of-sight limitations, especially in a crowded environment like the OR. The electromagnetic system provided a continuous stream of data. Their hybrid tracker employed a simple switching protocol: if the optical system is in view and available to collect data, it was used. If not, data was requested from the magnetic tracker. Only one piece of data was collected at each time interval, either optical or magnetic, so no true fusion was performed. This system was tested in an OR test set-up but not during an actual operation on a human. The main contribution of this group was to investigate to what extent ferrometallic materials in the OR affected a magnetic tracking system, and created a calibration look-up table to compensate for the interference. They also found that the calibration to be useful after multiple registration attempts under varying OR conditions. This is an idea that holds promise for future studies, and was not used in our study due to the fact that we could collect two separate data streams (optical and electromagnetic sensors) with relative ease. Creating a switching protocol would have been more time consuming.

Another group led by Nakamoto (2000) also created a hybrid system involving both optical and electromagnetic tracking systems. This group recognized that the source of many inaccuracies in a magnetic system in an operating room are the OR table and surgical instruments. Because of space and time constraints, it is also very difficult to calibrate for these distortions during or before an operation. This group developed a method for calibration, which allowed the magnetic transmitter to be moved intraoperatively, and allowed for optimal physical placement of the transmitter by using an optical sensor to track the magnetic transmitter. An interesting discovery by this group was that the distance between the magnetic transmitter and receiver must be relatively short to maintain an acceptable accuracy. They found that the transmitter-receiver distance must be 20cm for an error of 2mm in and around OR equipment. This group did not seem to fuse the data, but simply used the optical system to track the magnetic transmitter.

Our goal was to fuse the optical and magnetic data to create one continuous and high frequency dataset. This was the most reasonable and feasible option at the time as we were able to collect both the optical and magnetic data easily. We wanted to rely on the accuracy of the optical system, but use the continuous high frequency data from the magnetic system. By performing a data fusion, we were able to take advantage of the good qualities of both systems. The electromagnetic system we used was the Polhemus Fastrak.

The Polhemus Fastrak (Polhemus Inc., Colchester, VT, USA) electromagnetic tracking system was chosen to be the complementary tracking system to the Polaris optoelectronic tracking system. The Fastrak is a magnetically- based tracking system based on a fixed transmitter that sends out low frequency magnetic fields that allows the moving receiver to determine its position. Six degrees of freedom for position and orientation can be measured. The Fastrak does not suffer from line-of-sight issues, and has a much higher sampling frequency (120Hz). Although, magnetic systems do have their own disadvantages such as suffering from drift, interference from ferrous metals in the environment, and are electrically wired. The Polhemus user manual gives an accuracy of 2 mm within the 1m³ working volume, but one study found that this accuracy could only be achieved within a transmitter-receiver distance of 22cm (Milne 1996). The Polhemus Fastrak electromagnetic tracking system can be seen below in Figure 2.5.

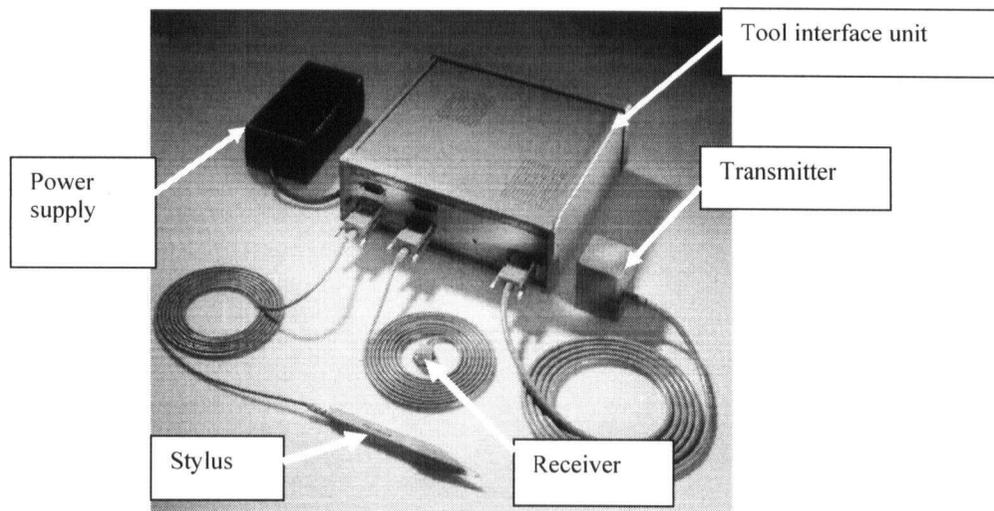


Figure 2.5: The Polhemus Fastrak magnetic position tracking system. This picture shows the tool interface unit, power supply, transmitter, receiver and the stylus.

2.3.2 Force/Torque

The adequate and appropriate use of forces/torques (F/T) in any surgical procedure is a skill that must be learned by a novice and practiced carefully by all. Surgical procedures require a certain amount of finesse and knowledge when applying forces and torques to human tissues. It is important that a surgeon is aware of this aspect, and takes it into account during any procedure. The collection of continuous high frequency F/T data to measure the surgical tool tip-tissue interaction force/torques during a live human surgery was our goal.

Rosen and colleagues have successfully measured forces and torques *in vivo* in a porcine model, and were able to classify surgeons' skill level using force/torque signatures (Rosen 1999). Their F/T data was collected using two separate sensors: a tri-axial F/T sensor and a strain gauge system mounted on the surgical tool handle (Figure 2.6). Their sensor is a custom-made tri-axial F/T transducer that mounts directly onto a laparoscopic tool shaft (hole through the center of the transducer).



Figure 2.6: *F/T system of Rosen (1999). A custom-designed F/T sensor was mounted directly onto the surgical tool shaft. This sensor has a hole through the center. A strain gauge system is mounted onto the tool handle.*

To measure the forces and torques associated with the surgical tool tip, we mounted a Mini40™ (ATI Industrial Automation, Apex, NC, USA) force/torque sensor (Figure 2.7) to our experimental Maryland dissector tool. This is a strain gauge based transducer, and able to withstand the forces and torques used in laparoscopic cholecystectomies. Force and torques in all three axes (F_x , F_y , F_z , T_x , T_y , T_z) were recorded at 120Hz in counts per unit force. This data is continuously collected directly into a Matlab file. Willem Atsma, also a member of the Neuromotor Control Laboratory, wrote the Matlab drivers for data streaming of the F/T data. The Mini40 was chosen as it was available in our lab, and is compact enough to be mounted onto the surgical tool without much interference, and not affect the weight of the surgical tool significantly.

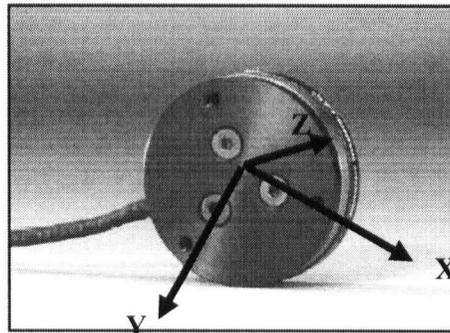


Figure 2.7: *ATI Mini40 F/T transducer. It is 40mm in diameter, 12.2mm thick, and weighs 50g. Sensing range: F_x , F_y +/- 80N, F_z +/-240N.*

2.3.3 Sensor Bracket Design

To attach the optical sensor MDMArray, magnetic receiver, and the F/T sensor, onto the Maryland dissector surgical tool, some type of mounting bracket was required. Many considerations were taken into account in the design of this mounting bracket (Table 2.1)

Table 2.1: *Criteria for design of the sensor mounting bracket.*

Criteria	Reason
Force bearing/load path	To allow for the force path to travel through the F/T sensor
Lightweight	To not affect the surgical tool weight and balance
Small	To keep the surgical tool shaft length as long as possible
Non-conductive material	To allow electro-cautery current to pass through the tool shaft and not through the sensors (especially the F/T sensor)
Non-obtrusive	To allow the surgeon as normal tool function as possible
No sharp edges	To prevent surgical staff from cutting gloves

Special measures had to be taken to allow the F/T sensor to be able to function properly, and be able to measure the tool tip forces/torques. We wanted to ensure that all the forces would be transmitted through the innermost shaft of the surgical tool. To do this, the outer shaft of the tool was cut to allow for these forces to be transmitted along the innermost shaft (Figure 2.8) through the bracket and then through the F/T sensor. This changed the original electrical isolation coating (seen as the thin black coating on the tool shaft) of the surgical tool shaft, and care had to be taken to minimize the area of the electrically live shaft that is exposed. The bracket was designed to not allow for accidental contact between the human and the exposed shaft.

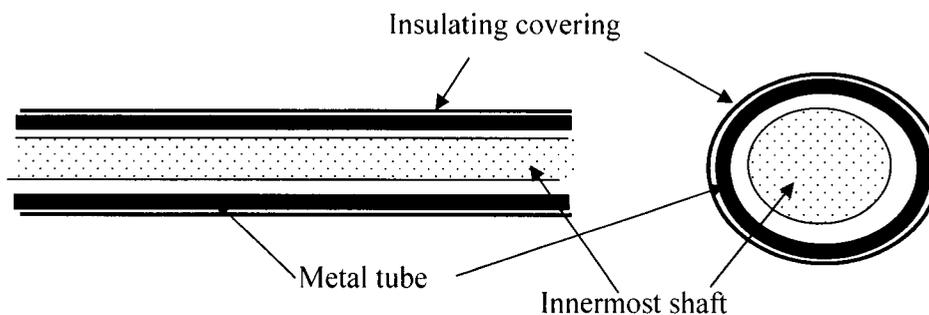


Figure 2.8: *Two cut views of a typical laparoscopic tool shaft. It consists of two layers of tubes (thin and thick lines), and the innermost shaft (dotted fill). The outer layer is a protective and electrically insulating covering. The middle tube is the metal structure of the tool shaft. And the innermost shaft is connected to the tool tip, and the tool handle. The innermost shaft and the middle tube are electrically live when electrocautery current is applied.*

After much iteration, the bracket was finally designed to mount all sensors, and satisfied the criteria (Figure 2.9). The bracket was designed in conjunction with volunteer lab engineer (Brandon Lee).

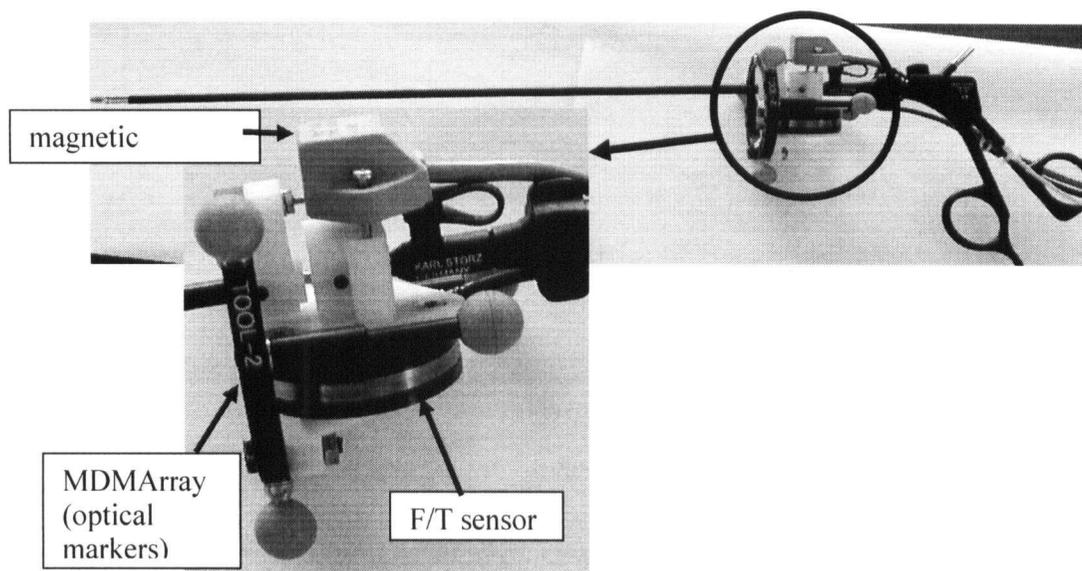


Figure 2.9: Sensor mounting bracket on surgical tool. The inset picture is a close up of the sensors mounted on bracket.

The final design of the bracket consists of two parts: top and bottom segments as seen in Figure 2.10. The bracket parts are mounted in-between the two parts of the original surgical tool. The top segment is directly attached to the outer shaft of the surgical tool, and will sense all forces of this top segment. The force sensor was not mounted inline, as with Rosen's device shown earlier, because our force sensor did not have a hole drilled through it to allow passage of the central rod.

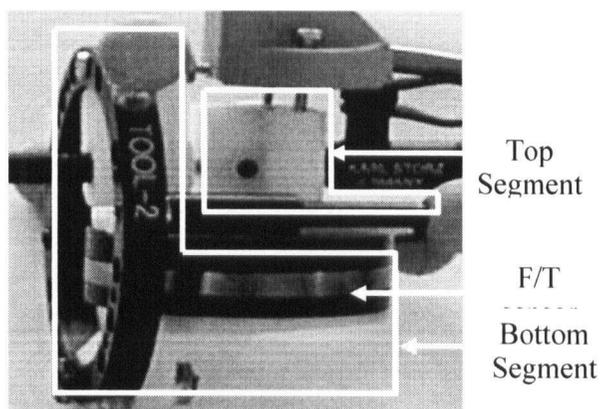


Figure 2.10: Force/Torque sensor bracket two segments. ATI force torque transducer mounted below tool shaft via custom designed bracket (Source: Kinnaird 2004).

The file was submitted in STL format (generated by SolidWorks) to technologists at the British Columbia Institute for Technology (BCIT) to construct the bracket out of a medical grade ABS plastic on the rapid prototyping machine. A non-conductive material was required because of the electrical current that is transmitted through the shaft for tissue cutting and coagulation. All the sensors, as well as the user and patient, must be protected from this electrical current. The design and material of the bracket also sets the magnetic receiver as far away as possible from any metallic elements that could potentially lead to errors in the magnetic sensor readings.

The wires coming from the Fastrak, F/T sensor, and the strain gauges, can all be gathered to one side of the bracket and tied together to minimize obstruction to the surgeons. This is done before each surgical experimental procedure. The detailed drawing and specifications of the mounting bracket can be found in Appendix C.

2.3.3.1 Force Balance

The MDMArray optical halo and F/T sensor connect the two segments. Since the force sensor is in the path connecting the segments, it registers any forces acting between them (Figure 2.11).

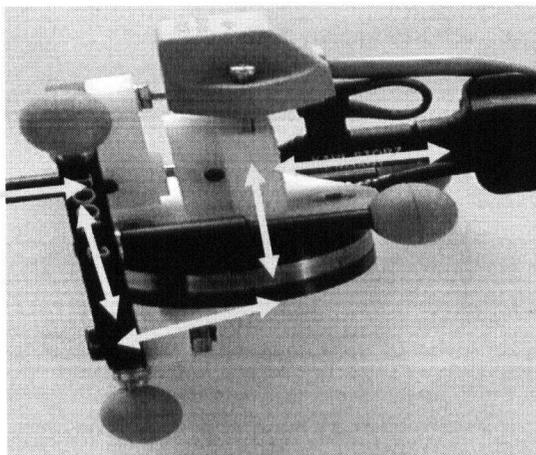


Figure 2.11: Force load path through sensor bracket and F/T sensor. The force travels bi-directionally along the tool shaft through the bottom segment, through sensor, and through top segment or vice versa.

In an OR situation, a trocar (tubular object used to hold surgical tool near operating site) is inserted into the abdomen, and the surgical tool is inserted through this trocar (Figure 2.12). This allows smoother movement of the tool and provides stability for the long laparoscopic tools. The surgical tool can be pushed down or pulled back along the length of the trocar to

access deeper tissues. The trocar also keeps the abdominal inflation gases inside because it is sealed. These gases are required in laparoscopic surgery to allow for better internal visualization.

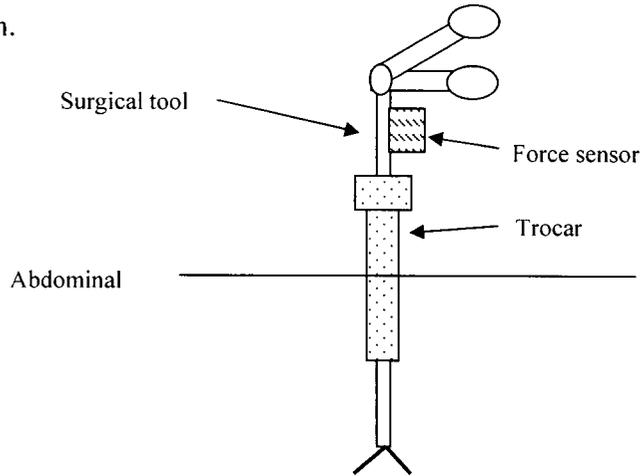


Figure 2.12: Laparoscopic trocar. The trocar provides stability for the long laparoscopic surgical tools. There are force interactions between the tool shaft and the trocar that are sensed by the force sensor.

In our subsequent data analysis, we require an estimate of the forces the surgeon is applying to the tissues using the tool. In this section, we present a free body diagram analysis of the loads applied to the tool and demonstrate how the tip forces are estimated. These FBDs are shown in Figures 2.13a-e.

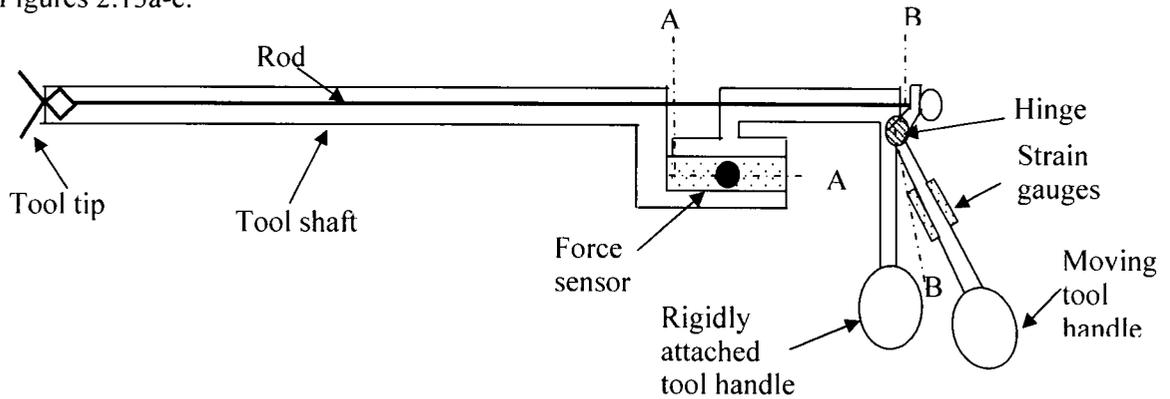


Figure 2.13a: Overall view of the tool, which is then split into 3 sections (2.13c,d,e) for FBD analysis. The dashed line at "A" represents the cut to create sections 2.13c and 2.13d. The "B" dashed line is used to create figures 2.13d and 2.13e.

In the following figures (2.13c-e), the following abbreviations are used. F_{ta} (actual tissue-tip interaction force), F_t (effective tissue-tip interaction force), F_a (force along the shaft, ie. trocar forces), F_r (tool rod force), F_s (sensor force), F_g (gravity force), F_h (hinge forces), and F_f (grip force of the hand on tool handles). The respective moments are also included.

The gravitational force (F_g) is assumed to be in the negative y-direction for illustrative purposes. In general, it will be a function of the tool's attitude. The effect of gravity forces on the force sensors is accounted for using a calibration method fully described in Kinnaird's thesis (Kinnaird 2004) and introduced in section 2.4 below. In the following FBDs, we identify the gravitational forces on the tool but in the subsequent analysis, we assume that the sensor readings have been adjusted to take these forces into account and therefore set the gravitational forces to zero.

There are two force-sensing elements in the tool; the force sensor collects forces and moments in all 3 directions (x, y and z) and the strain gauge pair is used to estimate the bending moment in the handle used to apply grasping forces. From these sensor readings, we are able to estimate the tip-tissue interaction forces as described below.

The F_t and M_t values are actually what we consider the effective tip force (ie. combination of both the actual tip-tissue interaction forces and any forces along the tool shaft). Unfortunately, due to the fact that the interactions between the surgical tool shaft and the trocar do not occur at a well defined point and are not directly sensed separately from the tip forces, the trocar interaction forces were not specifically modelled in this study. Directly estimating these trocar interaction forces would require a model that could account for the movement of the surgical tool along the trocar, but this is difficult because the trocar does not act on the tool at one specific point, but along a 7-10cm portion of the tool shaft. The characterization of the trocar-tool interaction forces could be investigated further in future studies. We did assume that the axial trocar forces were likely to not be very large in comparison with the tip/tissue interaction forces because the tool could slide through the trocar under its own weight. It is more difficult to justify a claim that the lateral forces are low because, although the abdominal wall is compliant and the tool is rarely used as a "pry bar", the forces could be comparable in magnitude. Nonetheless, since the point of application of the trocar forces changes, it is difficult to cleanly separate the two, which is why we have decided to represent the forces as equivalent tip forces and moments (i.e. tip forces + trocar forces = effective tip force), as shown in Figure 2.13b. Equations (a) – (f) show how the effective tip forces are affected by the presence of trocar forces.

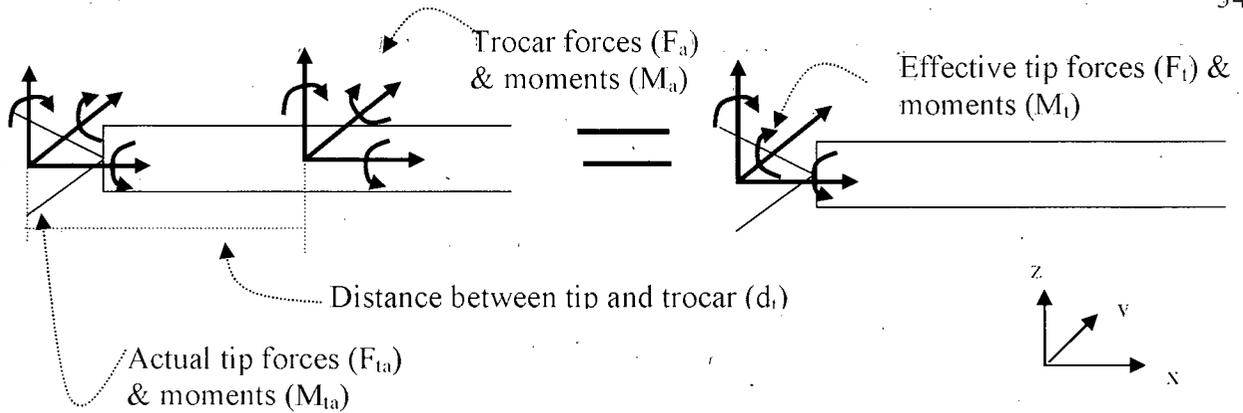


Figure 2.13b: Effective tip forces & moments are a combination of the trocar interaction force & moments and the actual tool-tip forces & moments.

The equations used to find the effective tip forces and moments are:

- | | |
|--------------------------------|--|
| a) $F_{tx} = F_{tax} + F_{ax}$ | d) $M_{tx} = M_{tax} + M_{ax}$ |
| b) $F_{ty} = F_{tay} + F_{ay}$ | e) $M_{ty} = M_{tay} + M_{ay} - F_{az}(d_t)$ |
| c) $F_{tz} = F_{taz} + F_{az}$ | f) $M_{tz} = M_{taz} + M_{az} + F_{az}(d_t)$ |

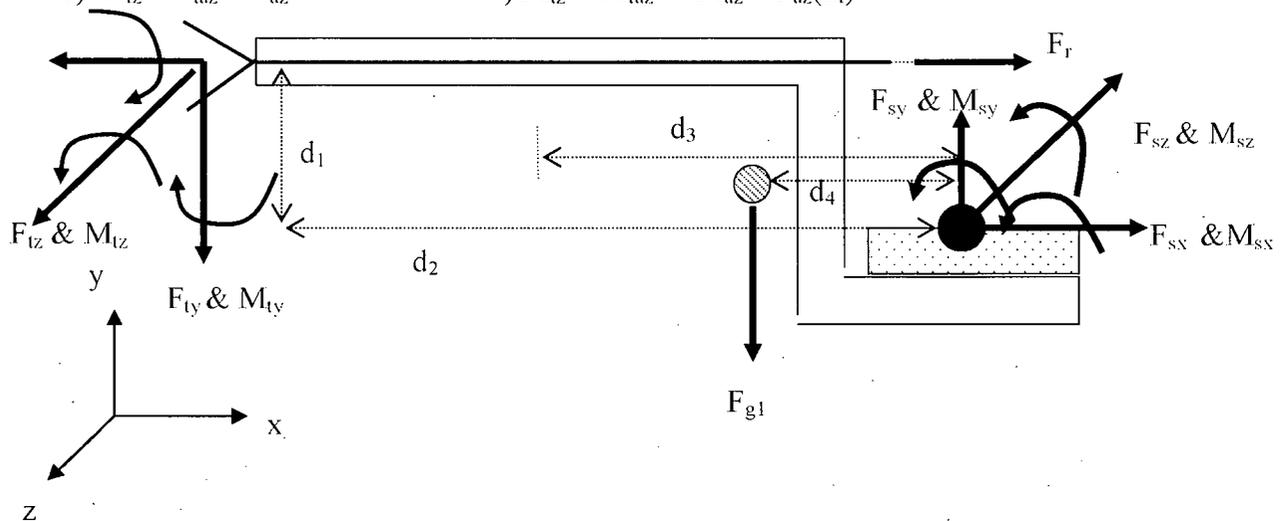


Figure 2.13c: Free body diagram of distal end of surgical tool and force sensor. Effective tip forces are represented. The "d" are the perpendicular distances of the forces used in the moment equation.

The equilibrium equations (assuming acceleration is comparatively low and can be neglected):

- 1) $\sum F_x = 0 = -F_{tx} + F_r + F_{sx}$
- 2) $\sum F_y = 0 = -F_{ty} - F_{g1} + F_{sy}$
- 3) $\sum F_z = 0 = -F_{tz} + F_{sz}$

Summing moments about the center of the force sensor (black circle on figure):

- 4) $\sum M_x = 0 = F_{tz}d_1 - M_{tx} + M_{sx}$
- 5) $\sum M_y = 0 = F_{tz}d_2 - M_{ty} + M_{sy}$

$$6) \quad \sum M_z = 0 = F_{fx}d_1 - F_r d_1 + F_{g1}d_4 + F_{fy}d_2 + M_{tz} + M_{sz}$$

Our goal here is to express the effective tip forces and moments in terms of the measured forces and the rod force.

- From eq.1: $F_{tx} = F_r + F_{sx}$ (eq.A)
- Applying gravity compensation ($F_{g1}=0$) to eq.2: $F_{ty} = F_{sy}$ (eq.B)
- From eq.3: $F_{tz} = -F_{sz}$ (eq.C)
- From eq.4 : $M_{tx} = F_{tz}d_1 + M_{sx}$ (eq.D)
- From eq.5 : $M_{ty} = F_{tz}d_2 + M_{sy}$ (eq.E)
- From eq.6 and gravity compensation: $M_{tz} = -F_{ty}d_2 + F_r d_1 - F_{tx}d_1 - M_{sz}$ (eq.F)
- Have 6 equations but 7 unknowns. The force sensor gives values for F_s and M_s ; F_r is derived from the analysis described later on page 36, which is found in Figure 2.13e.

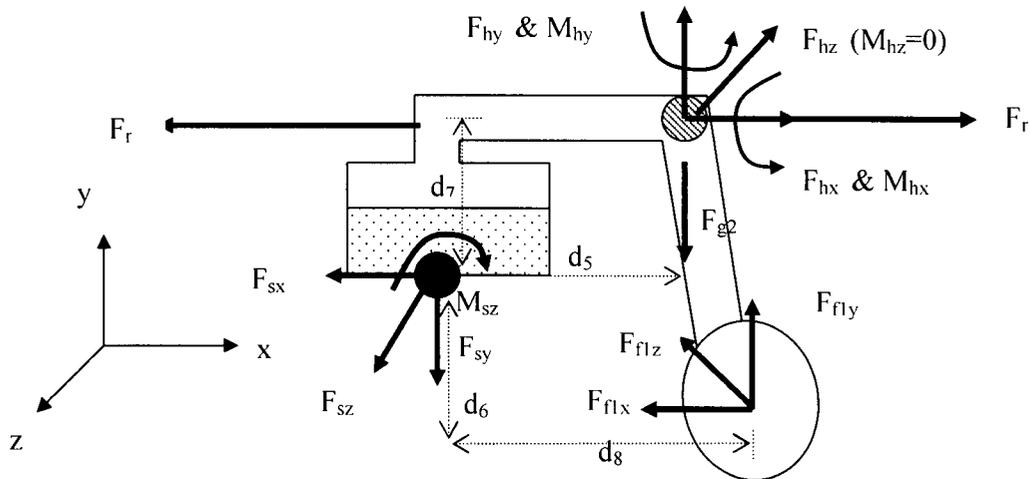


Figure 2.13d: Free body diagram of force sensor and stationary tool handle. The “d” are the perpendicular distances of the forces used in the moment equations, and are different for each section figure.

Note that this diagram is not used in the analysis, but is shown for completeness.

$$7) \quad \sum F_x = 0 = F_{hx} - F_{f1x} - F_{sx}$$

$$8) \quad \sum F_y = 0 = F_{f1y} - F_{g2} - F_{sy} + F_{hy}$$

$$9) \quad \sum F_z = 0 = F_{sz} + F_{hz} - F_{f1z}$$

Summing moments about the center of the force sensor (black circle on figure):

$$10) \quad \sum M_x = 0 = M_{sx} - M_{hx} + F_{hz}d_7 - F_{f1z}d_6$$

$$11) \quad \sum M_y = 0 = M_{sy} + F_{hz}d_5 - F_{f1z}d_8 - M_{hy}$$

$$12) \quad \sum M_z = 0 = F_{hy}d_5 - F_{g2}d_5 + F_{fy}d_8 - F_{fx}d_6 - F_{hx}d_7 - M_{sz}$$

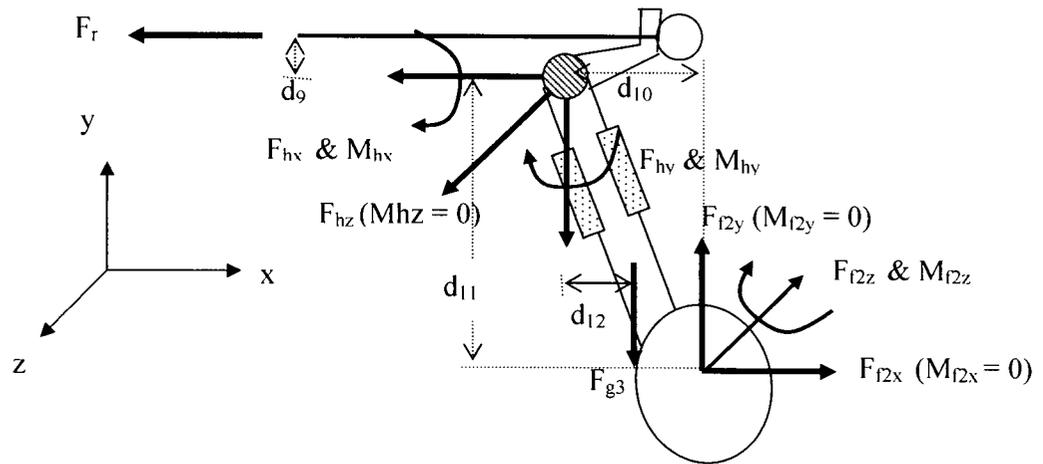


Figure 2.13e: Free body diagram of tool handle and strain gauges. The “d” are the perpendicular distances of the forces used in the moment equations, and are different for each section figure.

These equations are derived to show that the act of gripping results in the rod force, F_r .

$$13) \quad \sum F_x = 0 = -F_r - F_{hx} + F_{f2x}$$

$$14) \quad \sum F_y = 0 = -F_{g3} - F_{hy} + F_{f2y}$$

$$15) \quad \sum F_z = 0 = F_{hz} + F_{f2z}$$

Summing moments about the hinge:

$$16) \quad \sum M_x = 0 = M_{hx} - F_{f2z}d_{11}$$

$$17) \quad \sum M_y = 0 = M_{hy} + F_{f2z}d_{10}$$

$$18) \quad \sum M_z = 0 = F_r d_9 + F_{f2y}d_{10} + F_{f2x}d_{11} - F_{g3}d_{12} - M_{f2z}$$

- From eq.13: $F_r = F_{f2x} + F_{hx}$ (eq.G)
- Applying gravity compensation ($F_{g3} = 0$) to eq.14: $F_{f2y} = F_{hy}$ (eq.H)
- From eq.18 and gravity compensation: $F_r d_9 = -F_{f2y}d_{10} - F_{f2x}d_{11} + M_{f2z}$ (eq.I)

From eq. 18, if we make the assumption that F_{f2} (grip force of fingers) in Figure 2.13e is applied in the same fixed spot (as the finger holes for the tool are not large), and we take the sum of the moments around the hinge, we find that: $F_r = f(F_{f2}, M_{f2})$. Our strain gauge pair senses the bending moment in the handle at the gauge location (F_{f2})(grip to strain gauge distance) + M_{f2} . But we also believe that $M_{f2} \approx 0$ because it is physically difficult to apply a pure couple here. Therefore we can make the assumption that the strain gauge pair's output is

proportional to F_{r2} . So, in principle, we can compute F_r directly and substitute it back into eqs. 1–6 and the equations derived from them. In fact, it is more straightforward to observe the effect of grip forces on the force sensor output and to directly correct the force sensor readings as a function of the strain gauge pair's output, as described below in section 2.4; details are contained in Kinnaird's thesis (Kinnaird 2004).

Therefore, the final equations for estimating the effective tip force and moments are:

$$19) F_{tx} = F_r + F_{sx}$$

$$20) F_{ty} = F_{sy}$$

$$21) F_{tz} = -F_{sz}$$

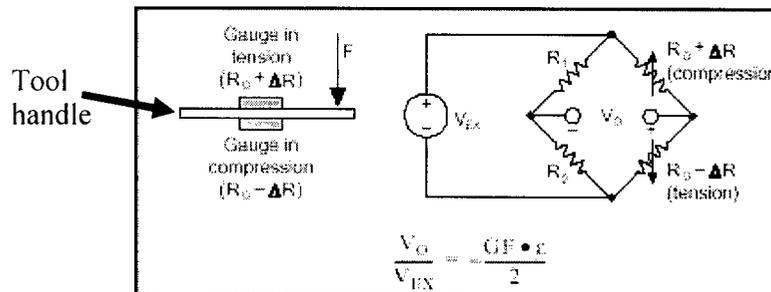
$$22) M_{tx} = F_{tz}d_1 + M_{sx}$$

$$23) M_{ty} = F_{tz}d_2 + M_{sy}$$

$$24) M_{tz} = -F_{ty}d_2 + F_r d_1 - F_{tx}d_1 - M_{sz}$$

2.3.4 Grip Force

The grip forces measured are used to correct the force readings to better estimate the tool shaft loads. The grip force is measured by two strain gauges mounted onto the surgical tool handle (Figure 2.14) and can the gauges be seen on the tool in Figure 2.15. In this half-bridge configuration, the gauges are measuring the perpendicular axis forces exerted on the handle, and we can correlate this force to forces at the surgical tool tip.



V_O = output voltage, V_{EX} = excitation voltage, GF = gauge factor, ϵ = strain
 R_G = nominal resistance of strain gauge, ΔR = strain induced change in resistance
 R_1 and R_2 reference resistors

Figure 2.14: Strain gauge circuit diagram for half-bridge configuration. The “ F ” represents the surgeon's grip force exerted on the tool handle.

The two gauges used were standard Vishay Micro-Measurements 120ohms. These two gauges are then fed into an instrumentation amplifier with built in gain and offset control. This signal conditioner also compensates for temperature by having the reference resistors within. The

These grip strains can then be extracted from the total F/T measurement to give a more accurate tool tip force measurement. This is explained further in section 2.4.1.

See Figure 2.15 for a picture of the strain gauges mounted to the surgical tool handle.

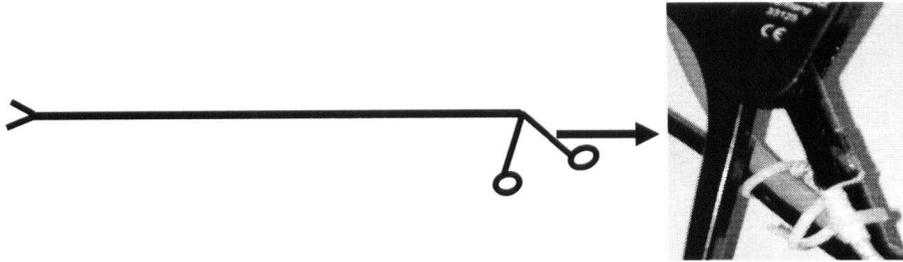


Figure 2.15: Strain gauges mounted on tool handle. The figure on the left is a general diagram of the surgical tool. The picture on the right shows a side view, where two gauges are attached on opposite sides of the tool handle.

2.4 Force/Grip Data Processing

In an earlier section, we showed that the force sensor also responded to grip forces. Here, we explain how we use the strain gauges to separate grip forces from tool interaction forces. The following sections describe our concerns with the force sensor calibration, and what was done to extract force data as a performance measure.

2.4.1 Grip Calibration

To fully understand and use the data received from the F/T sensor, an understanding of the mechanics of the surgical tool is needed (Figure 2.16). A typical laparoscopic surgical tool shaft has a few layers as described above in section 2.3.3. The innermost long shaft is what is attached to both the tool handles and the tool tip. This controls the opening and closing of the tool tip jaws. The tool handles are opened by the surgeon, the inner shaft will move and shorten, therefore causing the jaws to open due to the built-in pivot mechanism.

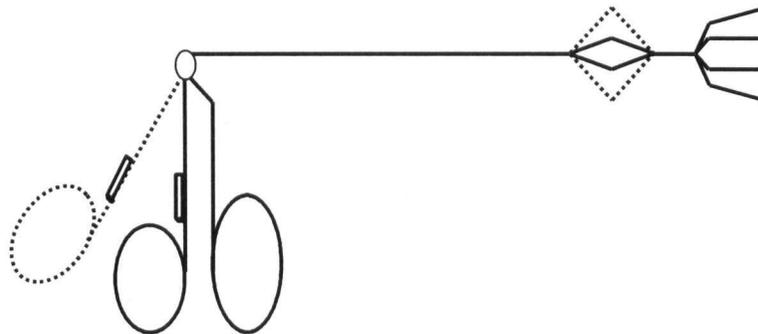


Figure 2.16: Mechanics of laparoscopic tool. When the handles are opened, the tool tip jaws are also opened due to a shortening of the innermost tool shaft. (Modified from source: Kinnaird 2004).

The F/T transducer senses this movement and will record it appropriately. Because of the design of our F/T sensor, the surgical tool shaft had to be cut, and the special bracket mounted as discussed previously in section 2.3.3. All the loads on the inner shaft are transferred through the bracket and sensed by the transducer. The interaction between the strain gauges and the force sensor is depicted in Figure 2.17. Through calibration, the strain gauge data is used to separate the grip forces from the actual tissue manipulation forces.

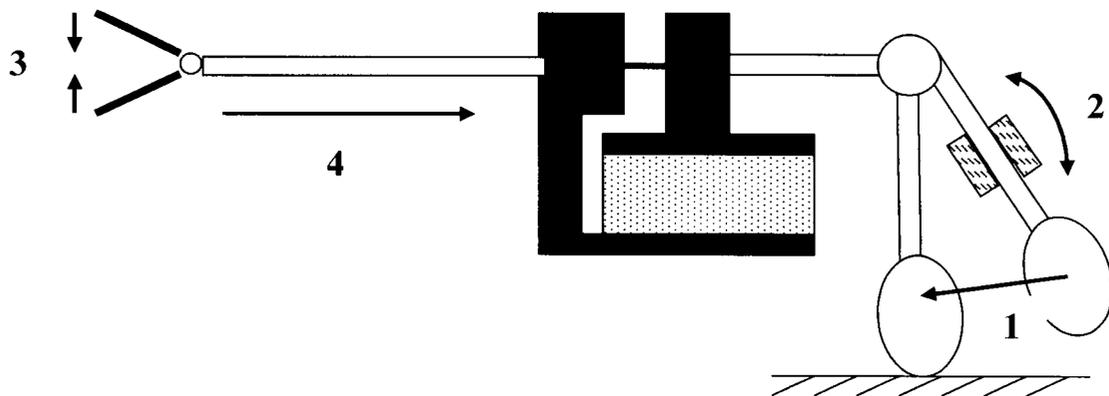


Figure 2.17: Interaction between strain gauges and force sensor. 1) Surgeon closes handle. 2) Strain gauges sense strain in tool handle. 3) Tool tip jaws close. 4) Tool shaft goes into compression, and the force sensor (dotted fill) senses this force against the tool bracket (solid black fill).

Discussion of the calibration algorithm used to separate the grip forces from tissue manipulation forces can be found in Kinnaird's thesis (2004). The tool was held in a neutral position, and the tool handle is open and shut while recording data from both the force sensor and the strain gauges. The results from one of Kinnaird's (2004) calibration tests are shown in Figure 2.18.

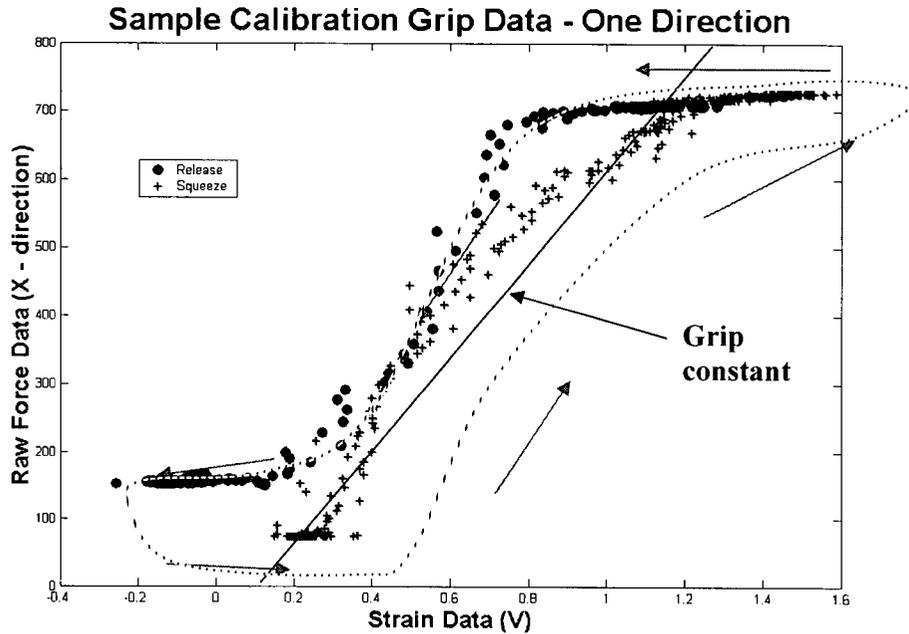


Figure 2.18: Results from one calibration test. The friction loop (dotted line) and the arrows are indicating the direction of motion. This loop is not consistent, and varies with grip strength. (Source: Kinnaird 2004)

Ideally, the force reading would be linearly related to grip strain, but there is clearly some nonlinearity. The force versus strain graph indicates flat sections at both ends where the strain reading increases while the force reading remains constant. The flat part at the lower left is likely due to friction within the tool handle (see Figure 2.19). The strain gauges detect the initial forces required to overcome the friction before any load is transmitted to the force sensor. There also seems to be a large amount of hysteresis during the release after the squeeze, as shown by the dashed line in Figure 2.18. This loop is not constant, and the location varies with different strength squeezes. The upper right portion of the plot also demonstrates another flat section due to saturation of the force sensor. The force sensor has overload protection but occasionally the surgeon's grip can reach the maximum sensing range of the sensor, which causes the strain reading to increase while the force reading stays constant. In conversation with Catherine Kinnaird, and a visual inspection of the force data verified that, the saturation problems (upper flat part of curve) were not very significant, as it is believed that less than 2% of the force readings hit this saturation area. This value was based on a visual inspection of the force data.

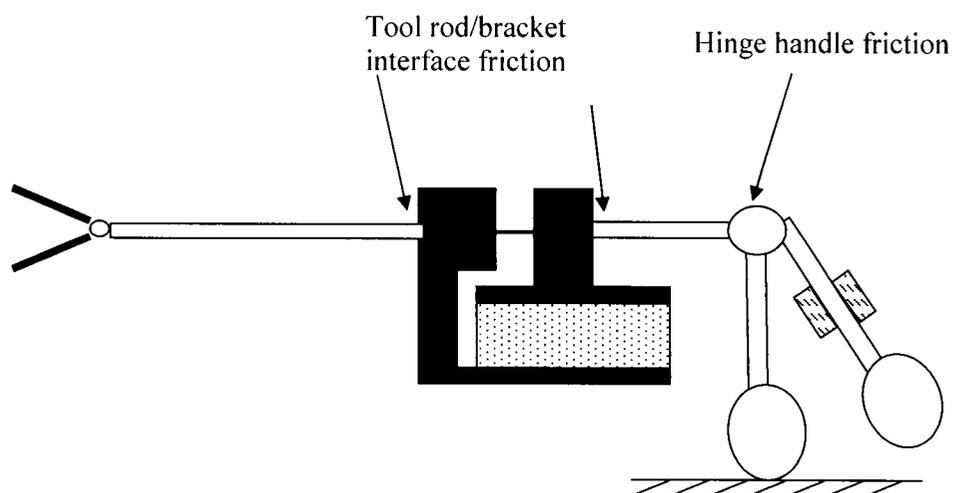


Figure 2.19: Friction in the surgical tool handle and bracket. There was friction in the tool hinge handle, and the tool rod/bracket interface. The mounting bracket also moved slightly/slipped along the tool shaft due to an improper fit. This movement may have contributed to the hysteretic loop in Figure 2.18, and to tool handle movements not being sensed by the force sensor.

These problems led to a need for a somewhat more complicated grip compensation algorithm than could be used if the relationship between strain gauge output and force sensor output did not exhibit either hysteresis nor saturation. A mean grip constant was calculated for the linear portions of the hysteretic loop (Figure 2.20). When the force data was within the linear range (inside the dotted oval), grip forces were removed. Outside of this range, no compensation was applied. On the upper side of this linear region, the force sensor is no longer responsive to changes in the tip force; no compensation was applied, leading to misleadingly high force peaks, especially in the axial direction. In conversation with Catherine Kinnaird, and by visual inspection of the force data, it was believed that about 20% of the force data might be outside the linear region (region outside of dotted oval). Kinnaird (2004) completed an error study of the compensation algorithm, and it was found that the algorithm reduces the RMS tip force error by about 50% in a typical manipulation (in a simulator environment). While it was correct to not apply any compensation to the force sensor reading when the grip force was in the low end of the curve, the correct thing to have done in the case of saturation would have been to set the tip force reading to “zero”, and if possible, to have excluded such data from subsequent analysis. However, we erroneously left the readings uncorrected, but believe this ultimately had a small effect on our conclusions because saturation occurred relatively infrequently.

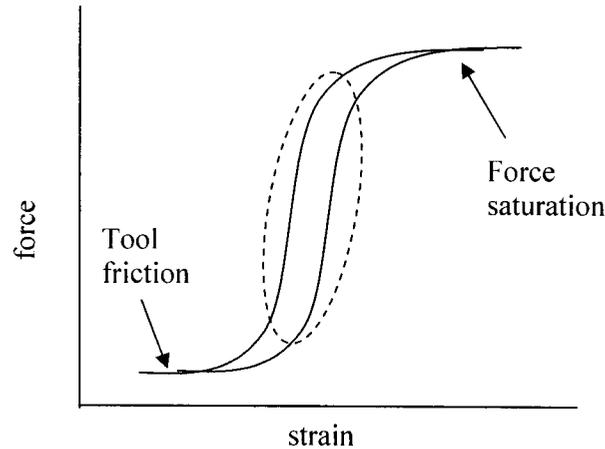


Figure 2.20: *Grip compensation. The force data has grip removed only in the linear region, as outlined in the dashed oval. This leads to misleadingly higher force peaks in the force data stream.*

With regards to tool design, the next iteration of the mounting bracket should be better fitted to the tool shaft to prevent the “slipping” movement mentioned above. The force sensor chosen should also be able to read the higher forces without saturating.

2.4.2 Gravity Effects Calibration

The effects of gravity on the F/T sensor are significant enough that they must also be compensated for. This sensor is inherently quite sensitive, and is mounted off the axis of the surgical tool shaft. This creates a force/torque reading just from the mass of the surgical tool alone. These F/T readings can vary up to $\sim 5\text{N}$ when the surgical tool is held or placed in different roll (rotation about z axis in tool tip frame (Figure 2.1)) and pitch (rotation about y axis in tool tip frame) orientations. Rotation in the yaw (about x axis in tool tip frame) does not affect F/T readings in the neutral position, as gravity naturally acts perpendicular to this axis of rotation. In order to improve the F/T readings, these effects of gravity must be compensated for.

A mathematical model was created to compensate for the roll and pitch. The details can be found in Catherine Kinnaird’s thesis (2004). This model led to an almost 2X decrease in RMS error when compared to a simple mean subtraction method, where the mean force value was subtracted from the total force.

2.5 Kinematics Data Fusion

Because we have used two different position sensors in our data collection process, we make use of this data for our kinematics measure, and have created a technique to fuse these two datasets.

2.5.1 Data Fusion Introduction

In a previous study in our lab by McBeth (2002), the Polaris optoelectronic tracking system was used to collect 3D position data collection for the method to quantitatively measure surgical performance. The optical data was collected at 20Hz in the McBeth study, which is a suitable sampling frequency for measuring human movement particularly for postural studies (Woltring 1986). However, as mentioned previously, one drawback of the optical sensor is that it is susceptible to line-of-sight problems, which leads to occlusion of the optical markers and consequently gaps in the optical data stream. He also suggested that a maximum gap size of 0.5s could be interpolated successfully (McBeth 2002). In the OR, marker occlusions longer than 0.5s occur quite frequently. These are the main reasons for wanting to improve the position tracking system for quantifying motor performance. Various different options were considered (i.e., ShapeTape, accelerometers, gyroscopes) as discussed in section 2.3.1.1.1, but in the end we chose to combine an electromagnetic position tracking system and the optical sensor.

Due to availability and ease-of-use, the Fastrak electromagnetic position tracking system was chosen to complement the Polaris. Fastrak is a three-dimensional (position and orientation) magnetic tracking system, and these are known to be free of line-of-sight issues, have continuous data collection, and sample at a high frequencies (~120Hz). According to the product manual, the 3D position and orientations of the receivers can be measured with an accuracy of 2mm and 0.15° within a 1m^3 working volume surrounding the magnetic transmitter, but we have found that in reality, the accuracy is not as good as the manual suggests. We have found that once the transmitter-receiver distance is greater than ~20-30cm, the data becomes less accurate, and tends to fluctuate around the actual value.

2.5.2 General Data Fusion

To obtain a useful estimate of the tool position, the two data streams must be fused. The general process of combining multiple data sources is well studied in a wide variety of

applications (Challa 2004). Sensor fusion is defined as “the combination of sensory data or data derived from sensory data such that the resulting information is in some sense better than would be possible when these sources were used individually” (Elmenreich 2002). In short, we would like to combine more than one source of data to create information that is better than either alone.

Table 2.2: *Advantages of data fusion (Elmenreich 2002)*

Advantage	Reason
Robustness & Reliability	Inherent redundancy and provide data even when one source fails
Extended Spatial & Temporal Coverage	One sensor can see where another cannot, and provide data when another cannot
Increased Confidence	Measurement of one sensor confirmed by measurements from other sensors of same domain
Reduced Ambiguity & Uncertainty	Joint information reduces set of ambiguous interpretations of measured value
Robustness Against Interference	Increasing the measurement space makes system less vulnerable to interference
Improved Resolution	Multiple independent measurements taken of same property

Sensor fusion can be implemented at various levels of interpretation depending on the application. Low-level fusion (or raw data fusion) will combine various sources of raw data to produce new data this is supposed to provide more information than the original inputs.

Intermediate-level fusion (or feature level fusion) combines features such as edges, corners, lines and textures into a feature map that is then used for segmentation and detection. High-level (or decision fusion) combines decisions from several experts. Methods include voting, fuzzy-logic and statistical methods. In our case, we will be concentrating on the low level or raw data fusion, as there will be two raw kinematics data sets combined into one.

Because the notion of data fusion covers so many levels of interpretation and types of sensors, there is no single model of fusion that can work for all applications. It is key to find a model that is optimal for a specific application.

2.5.2.1 Fusion Methods

There are many different types of sensor fusion, and each one has its pros and cons. Some of the areas where sensor fusion is used include the areas of military (i.e. target tracking), satellite positioning, and image processing (Challa 2004). The more common methods of sensor fusion in these areas include:

- Bayesian Inference – using probabilities to attach weightings. i.e., automotive applications sensor fusion (Coue 2003)
- Dempster-Shafter Inference – similar to Bayesian, but more computationally intensive. Allows for more unknowns, as it relies on “beliefs” and “masses”. i.e., New use in human computer interactions (Wu 2002)
- Artificial Neural Networks – perception studies (Johnson 1998)
- Kalman Filtering – prediction/correction filter, often used in navigation

2.5.4 Kinematics Data Fusion Technique

Before the collected data can be fused, it must be synchronized in time and registered into the same reference frame as summarized in Chapter 3 section 3.6.1. The details of synching and registering two data streams were presented in the complementary thesis of Kinnaird (2004). Once the two positional data sets are synchronized and registered, they are then put through the fusion process.

Position and orientation measurements from the optical sensor are considered to be correct and accurate when optical markers are visible, and the magnetic measurements are used to provide estimates of the shape and detail of the sensor’s trajectory, especially during times when the optical sensor has missing data (i.e. optical marker occlusions). We take these two data streams and fuse them into one continuous high frequency data stream. By using the accuracy of the optical data, and the continuity of the magnetic data, we take the advantages of both systems to get the data we want.

We first filter the magnetic data. This filtered magnetic data is then evaluated at the times of the optical data to estimate its value at the optical sampling times. The time matched optical and magnetic data is now subtracted from each other to create a difference curve. Next, we interpolate this difference curve to estimate the errors at each magnetic sample time. The fused data is an addition of the interpolated difference curve to the original magnetic data.

A flow diagram of the steps to data fusion is shown in Figure 2.21.

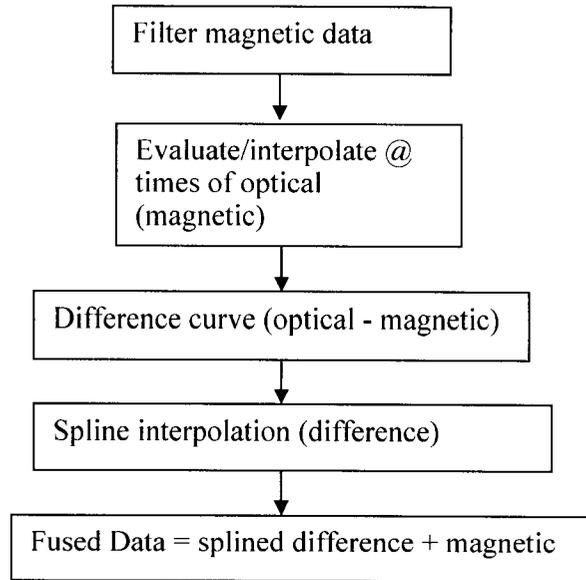


Figure 2.21: *Data fusion steps*

2.5.4.1 Data Fusion Technique Details

The original magnetic data was sampled at 120Hz, which is much faster than the 30Hz sampling rate of the optical data. The magnetic data was first filtered using a Generalized Cross Validation (GCV) approach by Woltring (1986) to smooth the dataset, as the magnetic data can be rather noisy (Figure 2.22a and Figure 2.22b).

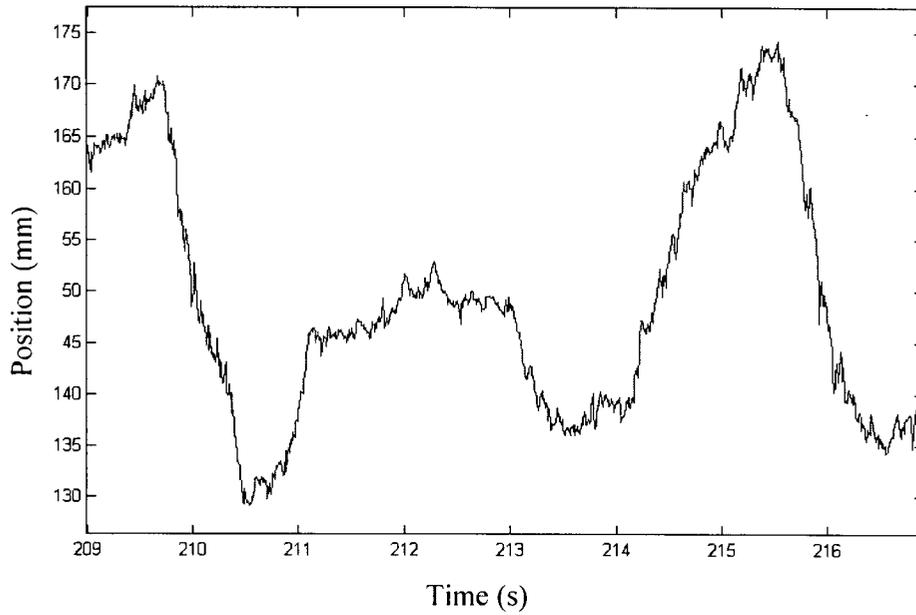


Figure 2.22a: *Noisy magnetic data. This noise is most likely caused by the environment (i.e. ferrous metals, medical instrumentation).*

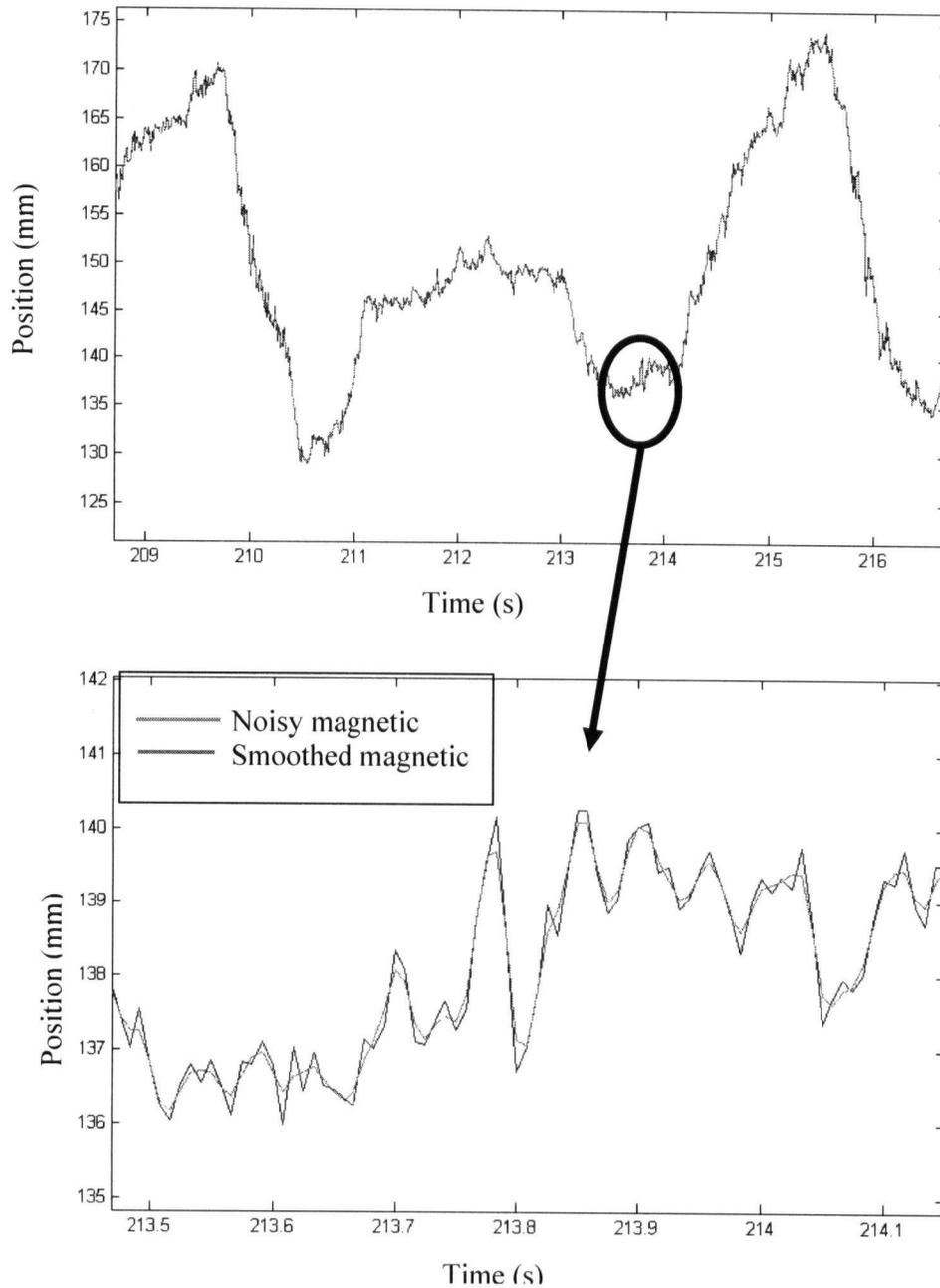


Figure 2.22b: *GCV smoothed magnetic data. The bottom graph is a magnification of the smoothed magnetic data.*

The algorithm iterates to find the optimal smoothing parameter by considering each data point and all the other data points to find a model that reproduces that one data point (i.e., minimum GCV is least affected by any single point). The GCV algorithm is of specific use in our application as it accommodates unequally time sampled data, and can handle multiple datasets.

We hypothesize that this noise is caused by the OR environment of metals and medical instrumentation systems, as the same amount of noise was seen in the laboratory setting. The noisy spikes of data caused by the electrosurgical unit (ESU) are also removed as will be described in chapter 3 section 3.7.1.1. The magnetic data points were then mathematically reduced to the same number of points as the optical data by down sampling to time match to the optical data. A difference curve was generated between the optical and interpolated magnetic data. This difference curve was then interpolated to estimate the errors at each magnetic sample time. The interpolated difference curve is then added to the original magnetic curve to produce the corrected/fused position estimate.

A demonstration of the data fusion technique is shown in the following figures. This data was collected in the laboratory to demonstrate typical operating room movements and the data fusion technique with typical data (Figure 2.23). One should also take note of the discrepancy in the registration of the magnetic and optical data (see Kinnaird's thesis 2004). This was one problem with the overall data registration system that our data fusion technique took care of, as will be demonstrated in the following sections.

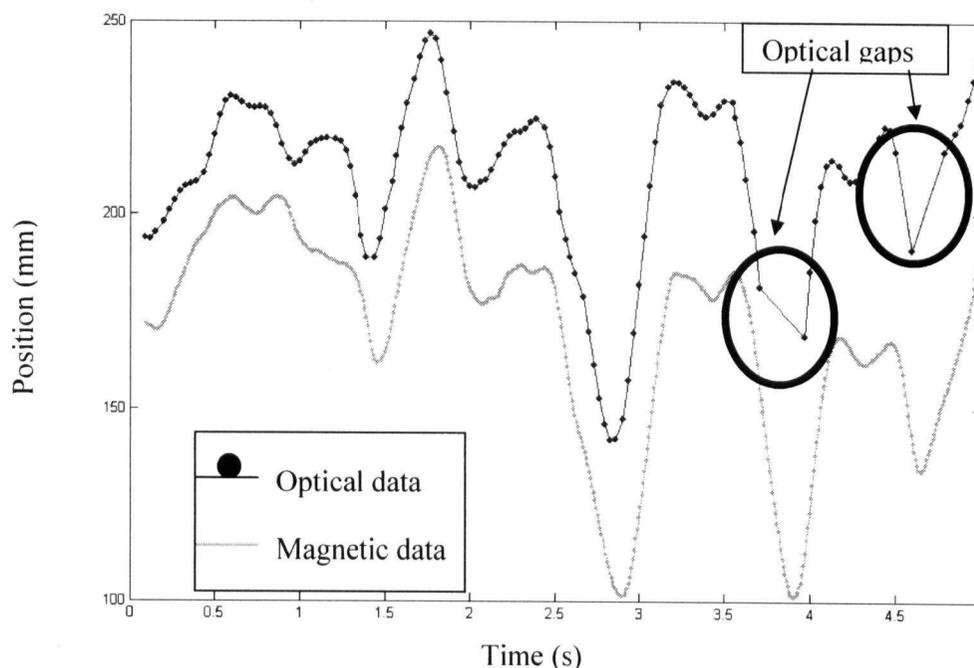


Figure 2.23: Laboratory data. Optical and magnetic data that has been time synched and registered. Optical gaps seen in circled areas.

The first step of GCV filtering the magnetic data was not done in this case, as the magnetic data was quite smooth, and not noisy. This is typical, as noise is usually seen more often in the operating room and rarely in the laboratory.

The next step is to interpolate the magnetic data to estimate its value at the optical sampling times (Figure 2.24).

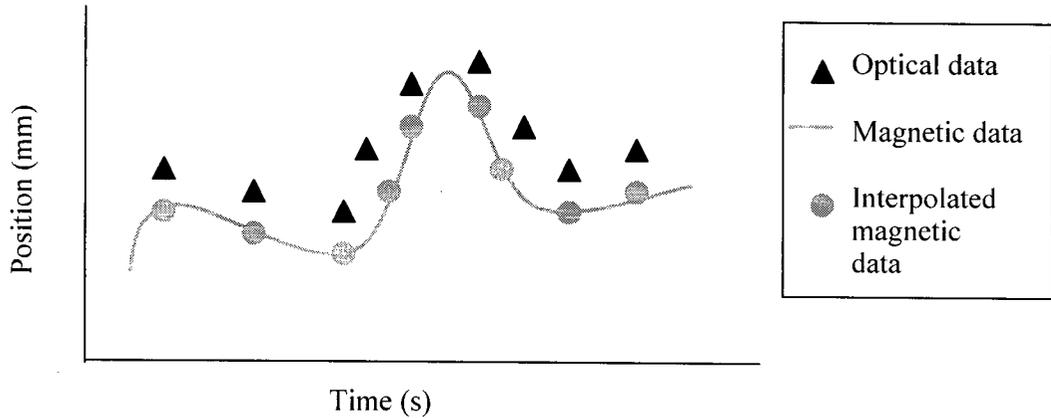


Figure 2.24: *Interpolate the magnetic data. The red magnetic data (line) is down sampled to time match the optical data that is sampled at a lower frequency (~30Hz). The red dots indicate the down-sampled magnetic data. This is simulated data.*

The Matlab function “interp1” is used to down-sample the magnetic data. The Matlab “interp1” is a one-dimensional data interpolation function based on a cubic spline algorithm. It acts like a lookup table to find the wanted data. This creates a magnetic data stream that is time matched with the optical data (Figure 2.25). Note that this process produces estimates only at times when optical data was available; if the optical sensor was occluded and a gap in the optical data stream resulted; there will be a corresponding gap in the down-sampled magnetic data stream.

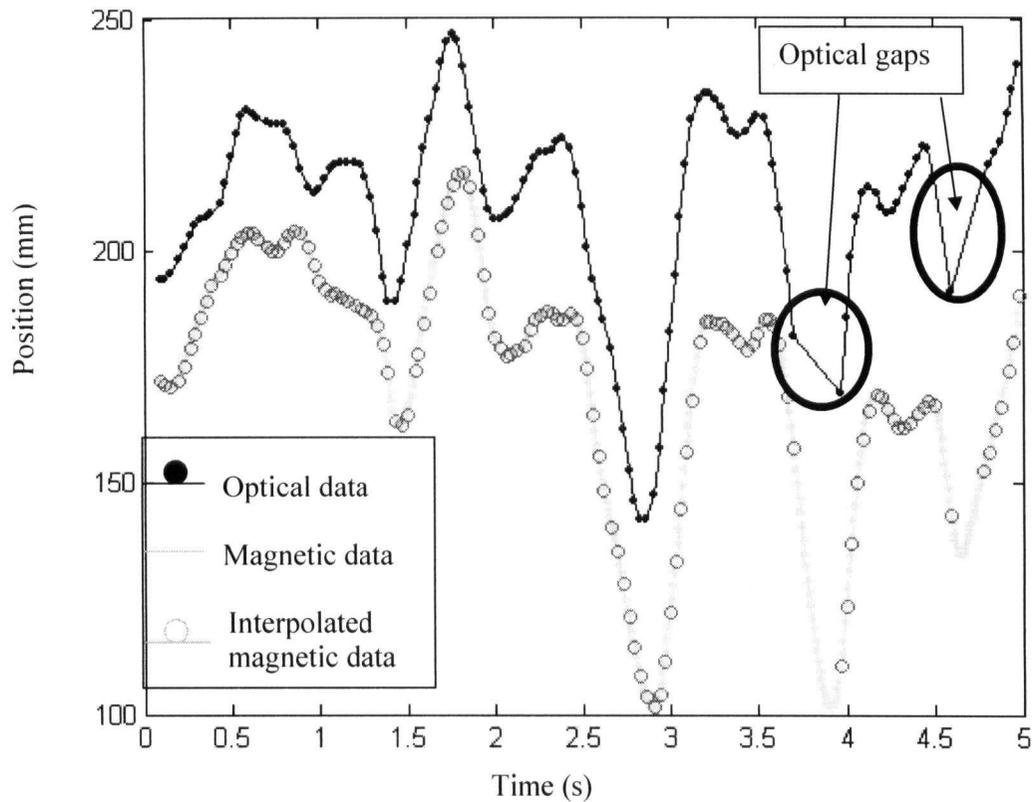


Figure 2.25: *Interpolated magnetic data. The line with open dots is the magnetic data evaluated at optical sample times. There are also gaps in the optical data (top line solid dots) as circled.*

The third step is to create a difference curve by subtracting the down-sampled magnetic data from the optical data (Figure 2.26). The difference curve represents the error in the magnetic data estimate at the available optical sample times.

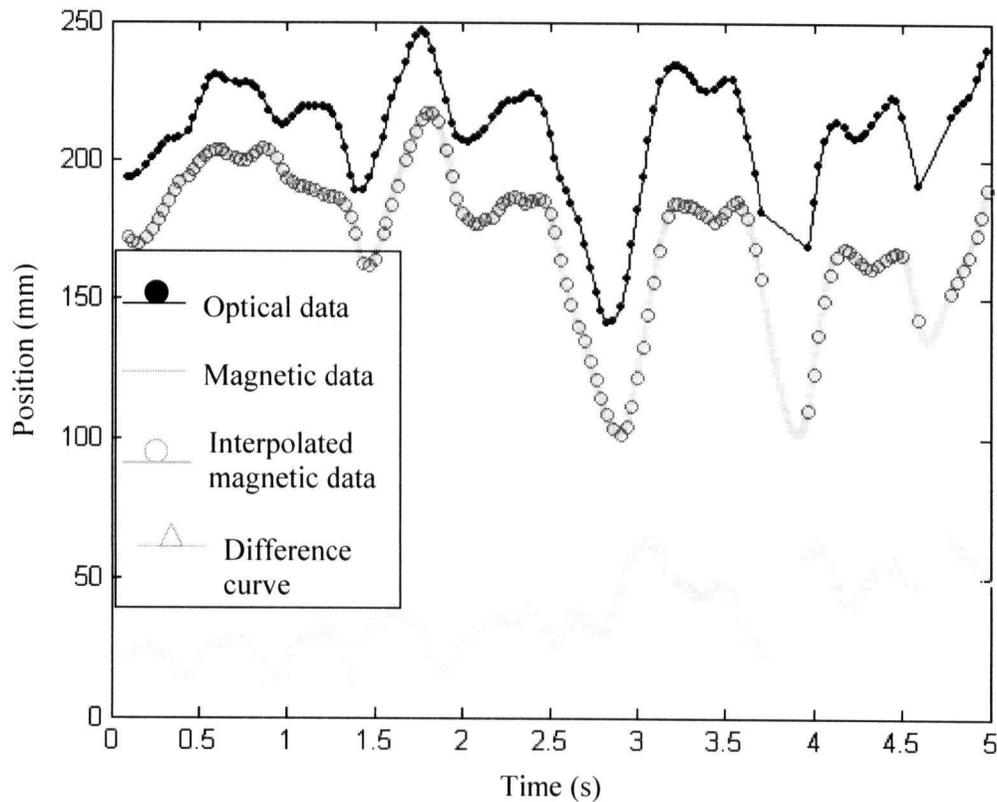


Figure 2.26: *Difference curve. The light blue line is the difference curve. This represents the error in the magnetic estimate.*

The fourth step is to interpolate the difference curve with the Matlab “interp1” function to estimate the errors at the original magnetic sample times (Figure 2.27). At times when optical data is missing (e.g. at ~3.7 - 4seconds), the error is estimated by interpolating the difference across the gap. If the magnetic and optical data are perfectly calibrated, registered and time-synchronized, then this difference curve will be constant. Any deviations from these assumptions will generally produce relatively low frequency and low magnitude deviations from this constant difference, so, in the absence of more specific information about how the difference curve varies in time, it is reasonable to simply bridge the gaps between optical fixes with a spline estimate. This process produces difference estimates not only across gaps in the optical data stream, but between sequential optical fixes as well.

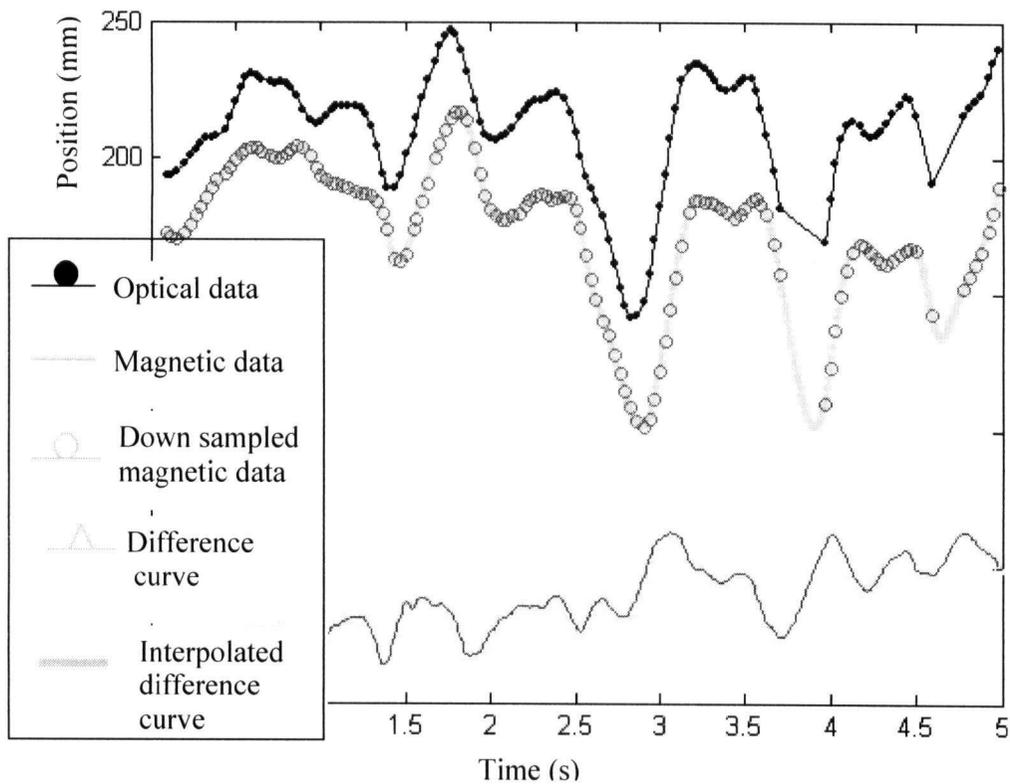


Figure 2.27: *Interpolated difference curve. The solid line is the interpolated difference curve.*

Finally, we create a by adding the interpolated difference curve back to the original magnetic data. This ‘fusion’ process treats the optical data as “fixes”, but it also fills in any optical gaps and produces a high frequency and continuous data stream at the sampling rate of the magnetic sensor (Figure 2.28). This effective increase in sampling rate aids in extracting performance measures of velocity, acceleration and jerk.

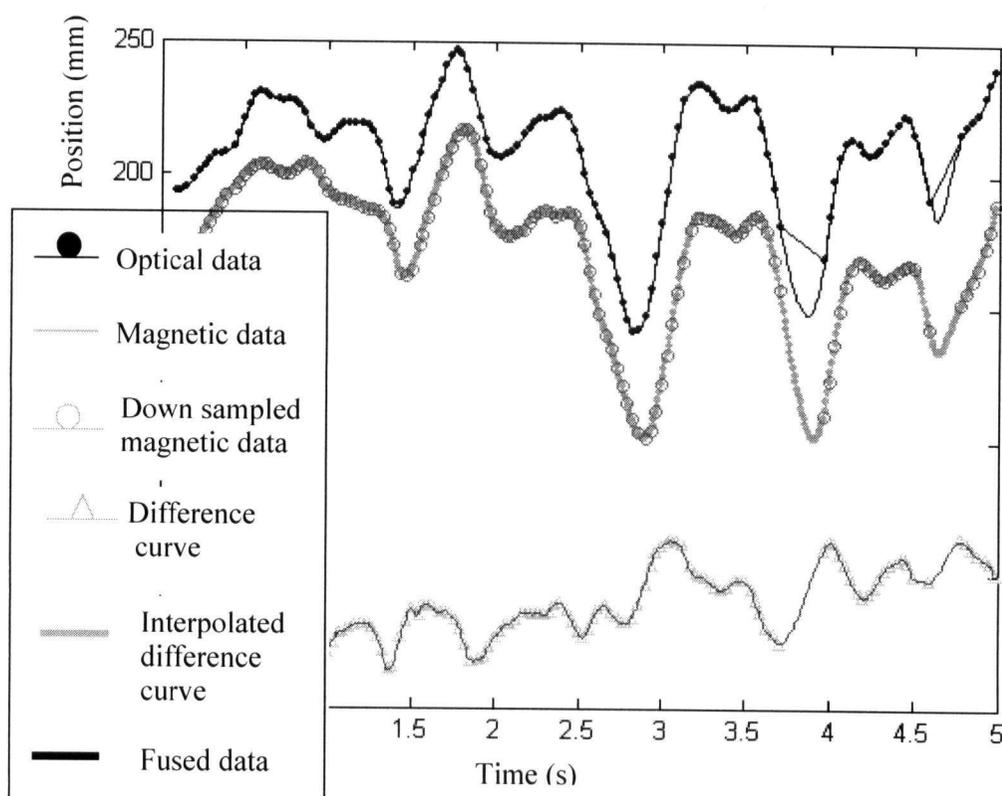


Figure 2.28: *Fused data. The solid black line through the optical data is the fused position estimate. It is a high frequency continuous dataset. Note how the data in the gaps is filled in based on the shape of the magnetic data and how the simple point-to-point connection of the optical data points across the data gaps produces significantly erroneous results.*

2.5.5 Error Analysis

To demonstrate the value of this data fusion technique, we compare the errors it produces to those of our previous optical interpolation technique using three sources of data: computer simulated data, laboratory collected data of typical surgical movements, and real OR data.

2.5.5.1 Analysis Method

The following technique was used for all three situations (computer simulated, laboratory collected, OR):

- 1) Collect a complete (without occlusions) optical data set for approximately 1-3 minutes.
- 2) Create artificial gaps in the optical data ranging from 0 - 10 seconds; after each application of the analysis, the gap is advanced by 0.1s and the calculations repeated.
- 3) Compute (a) interpolated optical (b) fused data sets.
- 4) Calculate RMS error across the gaps in both cases a & b

5) Compare a) and b).

The previously implemented optical interpolation algorithm was simply using the GCV algorithm to fill in the optical gaps. McBeth (2002) first chose the GCV parameters to be used for optical gap interpolation, and we used the same parameters.

2.5.5.2 Results of Error Analysis

Three sets of data were collected to conduct error analyses on. These were: 1) computer generated data, 2) data simulating typical surgical movements collected in the lab, and 3) actual OR collected data.

2.5.5.2.1 Computer Generated Data

The simulated data is a sine wave at a frequency of 1 Hz with an amplitude of 5 mm. These values were chosen because we felt they were of a similar frequency and amplitude to surgeon movements in the OR. The optical data is sampled at 30Hz, and the magnetic data at 120Hz, which are the same sampling frequencies that are used in the operating room experiments (Figure 2.29), and we add a 30mm offset to the magnetic data. This demonstrates the simplest form of our data fusion algorithm.

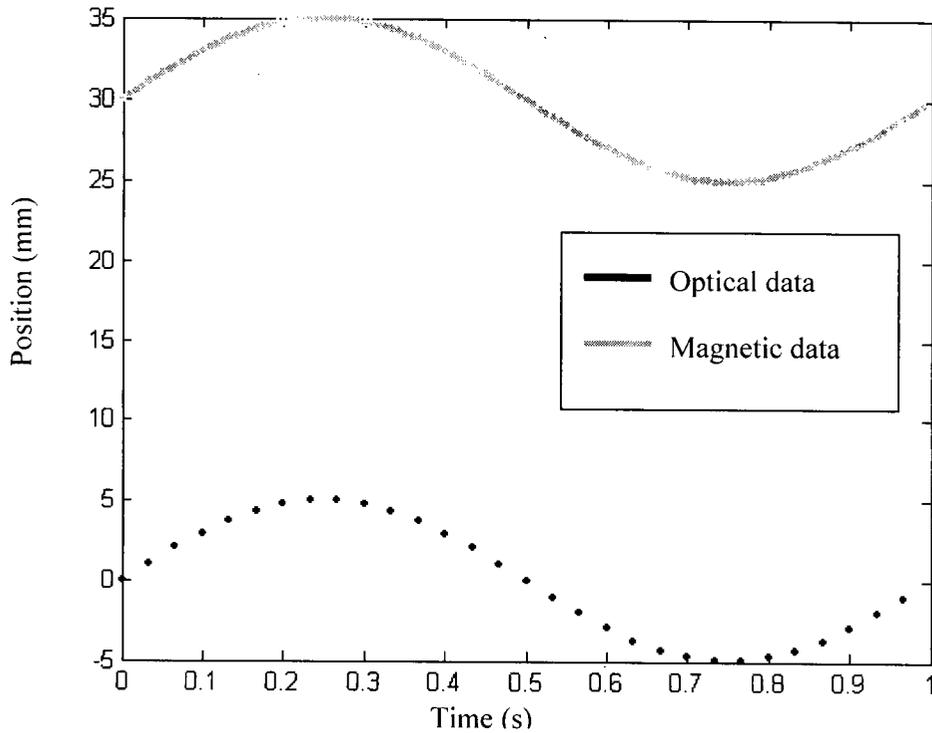


Figure 2.29: Computer-generated magnetic and optical data. The lower line is the optical data, while the upper line is the magnetic data. This plot only shows 1 second of data for better visualization of the sine curve.

RMS error analysis after using the interpolated optical and then using the data fusion technique is shown in Figure 2.30 below. The RMS error analysis involved taking the interpolated optical or fused data, and comparing it to the original. The differences between the original and interpolated or fused values are used in calculation of RMS error. The process was continued with larger gap sizes.

The optical interpolation method has much larger RMS error values as optical data gap size increases. The data fusion technique error is so small that it is negligible, and is shown as zero on the plot (RMS error = negligible). In reality (i.e., in the OR), the amplitudes of the magnetic and optical data would not match precisely, so this performance is unlikely, but it illustrates the concept.

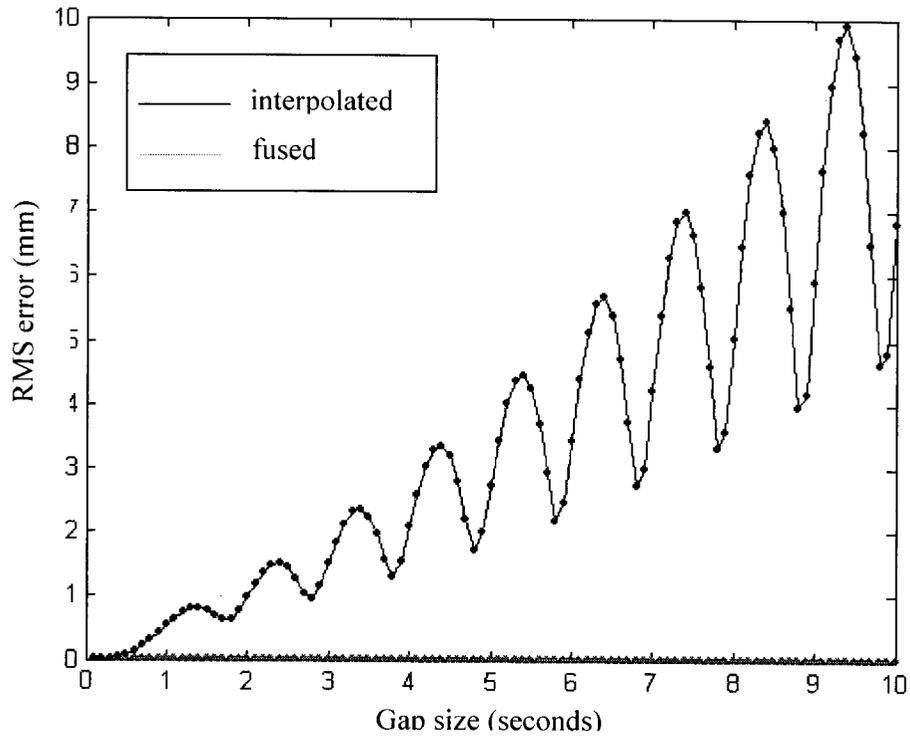


Figure 2.30: *RMS error for computer-generated data.*

2.5.5.2.2 Laboratory Data

Data was collected in the laboratory using movements similar to those that would be performed in the operating room. This would give us an idea of how our data fusion technique would work with typical data, but without having to set up and collect data during a live operation. We would not have to deal with electrosurgery unit (ESU) effects or other general noise created by the OR environment. Almost 70 seconds of data was collected for error analysis (Figure 2.31).

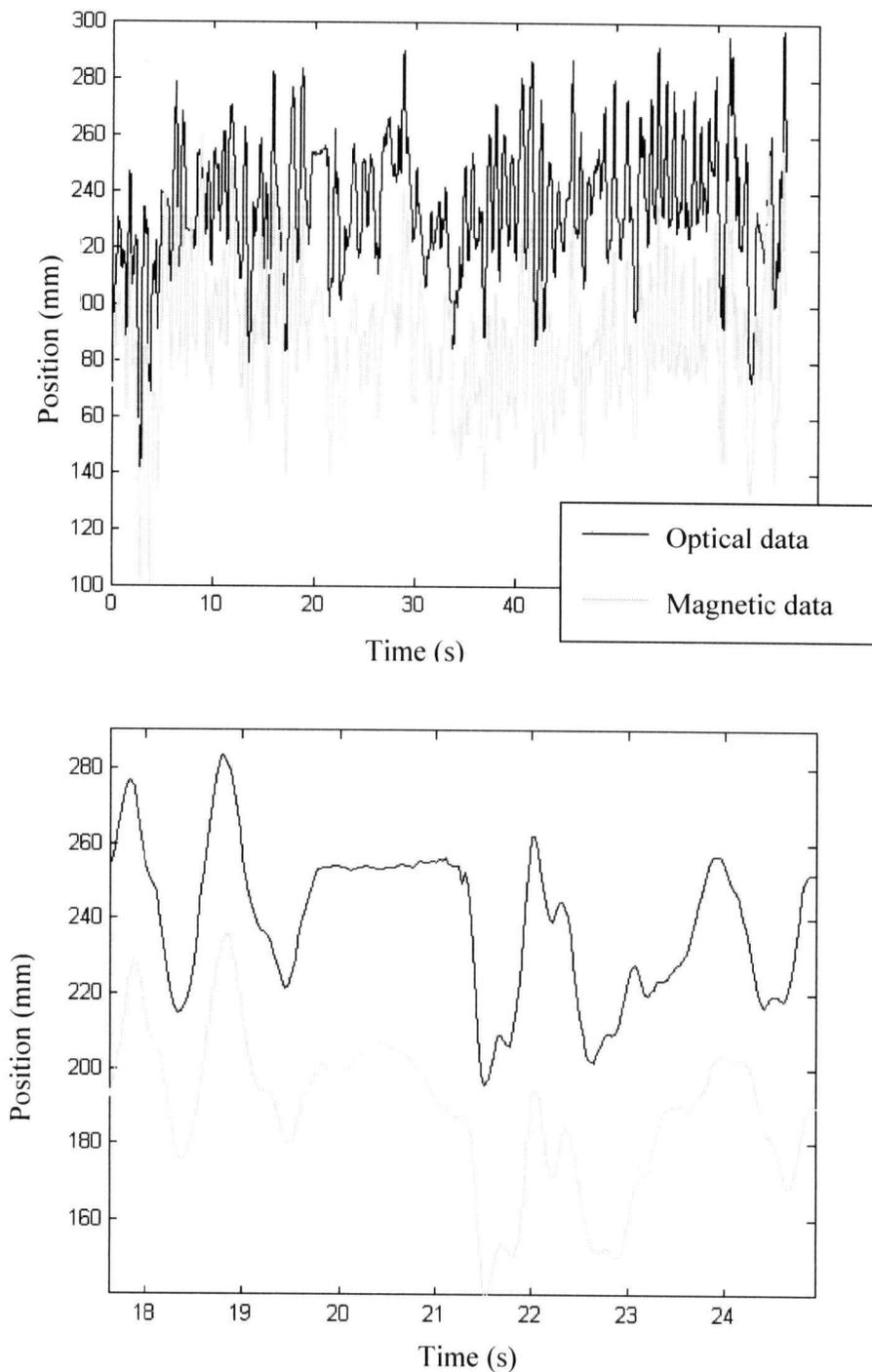


Figure 2.31: Laboratory collected magnetic and optical data. Bottom figure is a closer view of a 6-7 second interval.

Again, we looked at using only the interpolated optical technique compared to our data fusion technique. The results again show a large improvement in the RMS error. This reduction in

error is significant as the data fusion process filled in the gaps of the optical data, and was especially effective across the larger gap sizes (Figure 2.32).

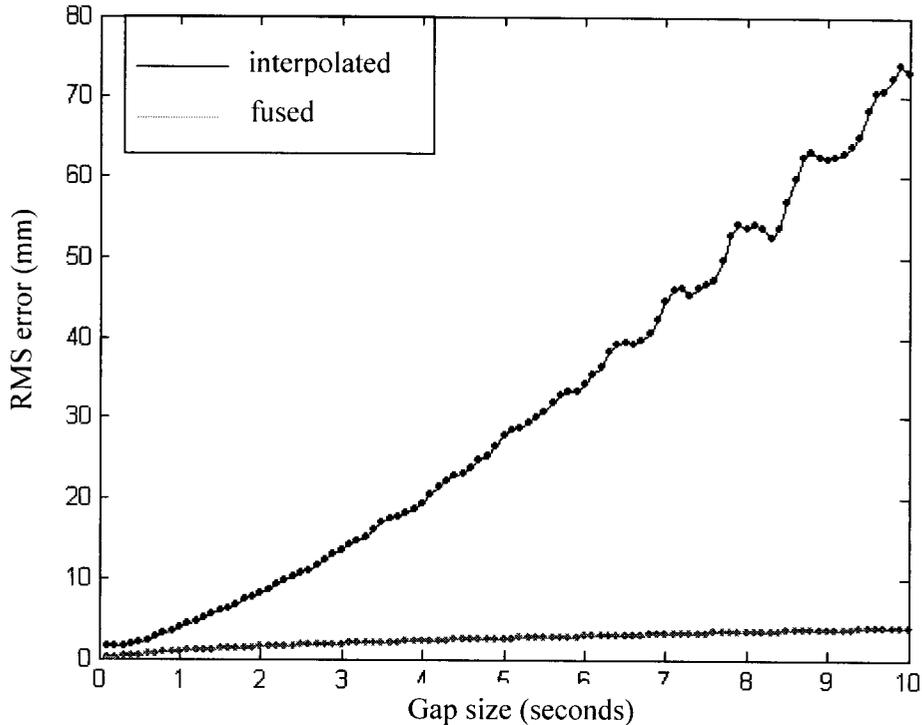


Figure 2.32: *RMS error for laboratory collected data.*

We also note that the error of the fused data in the laboratory data is below 5mm RMS error, even with a gap size of 10 seconds. At a gap size of 10 seconds, the fused RMS error is 3.9mm, while the interpolated error is at 73.2mm. This demonstrates that the magnetic sensor is able to capture the high-frequency variations in the position signal across the gap, whereas the simple interpolation algorithm essentially assumes a simple spline shape across the gap, thereby producing significant error.

2.5.5.2.3 Operating Room Data

Operating room data was extracted from one of our experiments to see how well the data fusion technique worked with real OR data. Approximately 200 seconds of data was extracted for error analysis (Figure 2.33).

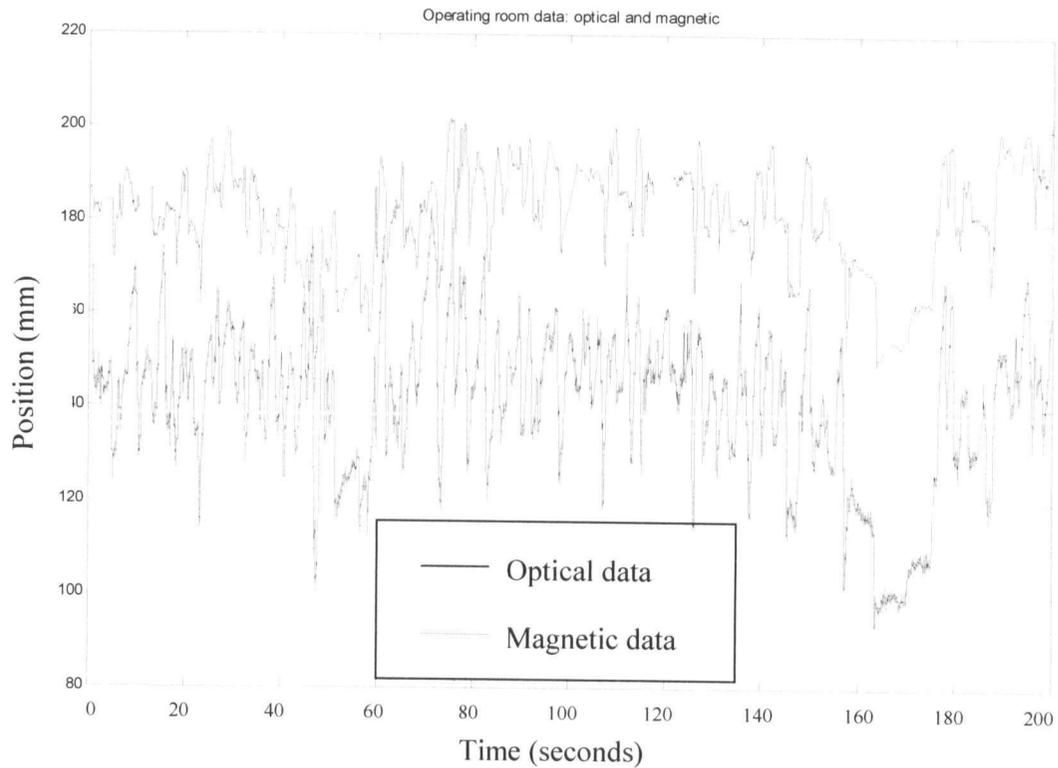


Figure 2.33: *Real OR magnetic and optical data*

We see that the RMS error is again lower with our data fusion technique when compared to the interpolated optical (Figure 2.34). The fused error at the 10second gap size was 2.38mm, while the interpolated data error was 28.15mm. There is a large improvement in RMS error with use of the data fusion technique.

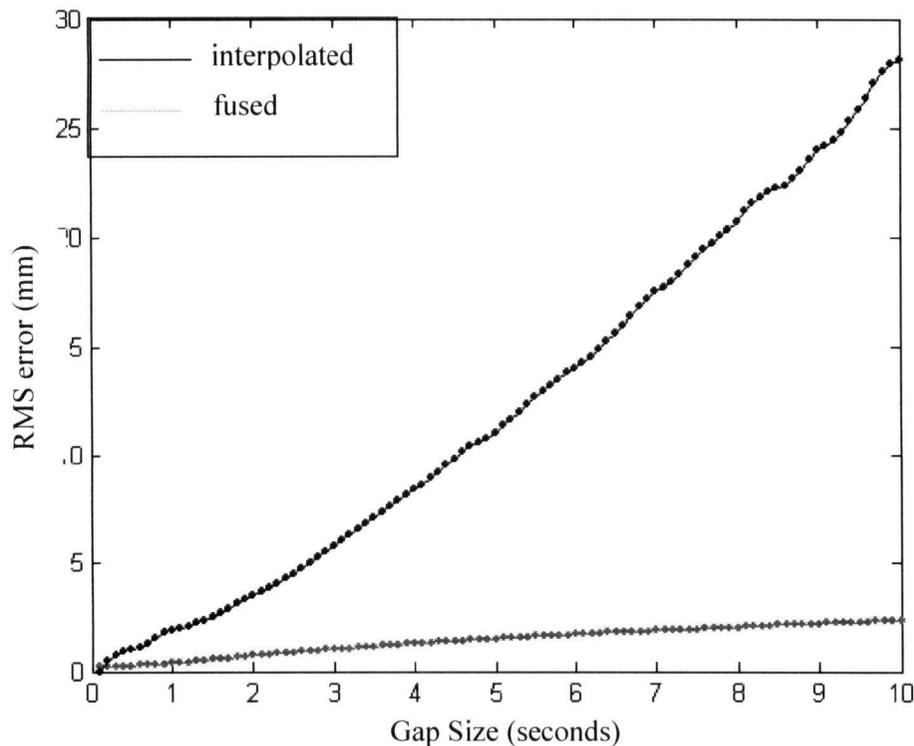


Figure 2.34: *RMS error for OR data.*

2.5.6 Discussion of Kinematics Data Fusion

Although, in our experience, we did not get gaps of over 10s in our real OR data, a check was done to see what happened to the RMS error at gap sizes of 20s and 30s. The RMS error with the fused data continued to increase slowly, while the RMS error of the interpolated optical signal also continued to rise, but at a much steeper slope. The optical error is related to the magnitude of tool excursion, and the magnetic/fused error is a function of the intrinsic accuracy of the magnetic sensor. Therefore, by fusing the data from both sensors, we are able to come up with an accurate and high-frequency estimate of the tool's position.

For typical OR movements in the range of 20mm, the kinematics fusion algorithm can give us a 10-fold decrease in RMS error. In the OR, we also see much finer positional movements (i.e., small millimeter movements) as compared to our other settings. And we see that with large movement excursions (50mm), as seen in the laboratory data, we get even a larger reduction in error with the use of our data fusion algorithm.

We specifically conducted our error analysis of data fusion for translations, and this shows good results. Specific analysis was not done for rotational data.

2.5.7 Conclusions for Kinematics Data Fusion

For our experiments, the final objective was to obtain high frequency, continuous estimates of velocity, acceleration and jerk of the surgical tool tip. We recorded data from both the optical and magnetic systems and applied a novel data fusion technique.

By using our data fusion technique, we have collected a more complete, high frequency continuous data set. This is much improved over the past technique used in our lab of only using optical tracking systems. This data fusion technique also compensated for discrepancies in the registration of data (as discussed in Kinnaird's thesis 2004).

This newly proposed data fusion technique is simple and effective for the purpose of fusing two data streams of similar position and orientation data. It allows for the combination of an optical and magnetic tracking system that improves the error over interpolation of the optical data alone. Because of missing optical data due to loss of sight of the markers, this led to difficulties in calculating derivatives of the position data. By using the magnetic tracking system that has uninterrupted data collection, and warping it to the optical data, we have solved our issues of missing optical data. The interpolation of the optical data was sufficient for the previous studies, but the data fusion technique created produces vast improvements over the interpolation method. As with most data fusion techniques, it is hard to pick one fusion technique that will be optimal for all situations. But for our situation, this new method seems to work very well. For its simplicity and novelty, our data fusion technique meets our objective to create a high frequency data set.

2.6 Discussion and Recommendations

The purpose of this project was to create a system that would be able to collect quantitative performance measures so we could objectively assess the construct and performance validity of both physical and virtual reality simulators. This involved modifying an existing system to be capable of collecting high frequency continuous kinematics and force/torque data from a laparoscopic tool during a live human surgery.

One contribution was the development of a system that was able to collect continuous high-frequency kinematics and F/T data in a live human OR with minimal disturbance. The use of a combination of sensors including optical and magnetic position sensors, F/T transducer, and strain gauges all allowed us to achieve our goal. The fusion of the optical and magnetic tracking systems was a novel method to overcome the previous problems of marker occlusion with the optical system alone. By joining these two position-tracking systems, we were able to have the positive features of both sensors: accuracy of the optical system, and high frequency and continuity of the magnetic system.

There are recommendations suggested for the continuation of this work in the future, as well as what steps could be taken to improve the overall system in various ways.

2.6.1 Kinematics

The kinematics system has been in use in our lab for the past 5 years, and we have continuously improved this system to gather reliable and accurate kinematics measures from the human OR. Because of the optical data gaps caused by occlusions of the markers in the OR, we have chosen to include a second position tracking system. The two position tracking systems (optical and electromagnetic) data streams are fused together to create a high frequency continuous data set. The new kinematics fusion technique demonstrates a large improvement over the previously implemented optical interpolation technique to fill in the optical gaps.

2.6.2 Force

The force measures are a new addition to our performance measure data collection system. We have created a system involving both a force sensor and strain gauges to analyze tissue interaction forces. This system is the first system that is able to collect tool forces in the human OR during MIS. But there are some revisions that need to be made to derive more accurate force measures.

The grip calibration scheme should be revised. The grip compensation was more complex than originally thought (problems with surgical tool friction and force sensor saturation). The algorithm could not compensate for all forces, and may have led to misleadingly high forces.

Rosen (1999) published their work with similar raw force data as was found in our study. They were also not able to properly distinguish the grip force from the force sensor data. In theory, the idea works quite well, but we believe that with a few improvements to the overall system, better force estimates could be made. A redesign of the mounting bracket, and a force sensor that could be mounted directly onto the tool shaft could possibly help to eliminate issues such as the movement of the bracket on the tool shaft.

Dr. Blake Hannaford (2004) of the University of Washington mentioned [personal communication] that he believes almost 80% of the forces sensed by the force transducer come from the trocar interaction with the surgical tool. We believe that it may be possible that lateral forces could contribute that amount, but axial forces at the trocar would not be as significant as the surgical tool does not experience much axial friction going through the trocar. Dr. Hannaford's group has done extensive study in this area, as they were one of the first groups to mount a force sensor onto a surgical tool (Rosen 1999). This is a possible significant contribution to the force measures, and could be taken into account. But on the other hand, we have measured all tool interaction forces, regardless of the source, and this in itself is an interesting measure. Laboratory studies could be conducted and models created to determine exactly how much force is created by the interaction between the surgical tool shaft and the trocar. Compensation algorithms could then take these trocar forces into account to reveal more accurate tool-tissue interaction forces.

2.6.3 Recommendations

The data collection and analysis system that we have developed for the quantitative assessment of surgical performance is a unique system, and is one of the first to be used for this kind of data collection in the human operating room. Because this system is a first attempt at such a difficult endeavour, there are some improvements that could be made for even easier data collection and analysis. The recommendations include:

- Sensor mounting bracket:
 - Location of each of the sensors currently changes the “weighting” of the normal surgical tool causing unnatural rotation of the tool tip. The bracket seems to be

“bottom heavy”, and will want to swing into the one position. Rearrangement of where each of the sensors is mounted may help with this issue.

- Possible friction issues where bracket meets the inner tool shaft causing “stickiness” in the movement of the tool handles.
- Flex and warping of the actual sensor mounting bracket. The material should be stiffer.
- Current position of the magnetic receiver occludes the optical markers. Bracket redesign or repositioning of the receiver would help.
- Force/Torque sensor:
 - Physically smaller sensor would take up less space on the mounting bracket. And would also help with the “weighting” issue.
- Data acquisition:
 - One button operation from one computer to alleviate some problems with space and time in the operating room.
 - Possible use of different software system that is designed for data acquisition with multiple serial ports.

The experimental data collection and analysis system allowed the collection of data in the human operating room. With few improvements, the system could be made for easier widespread use in larger studies.

Chapter 3

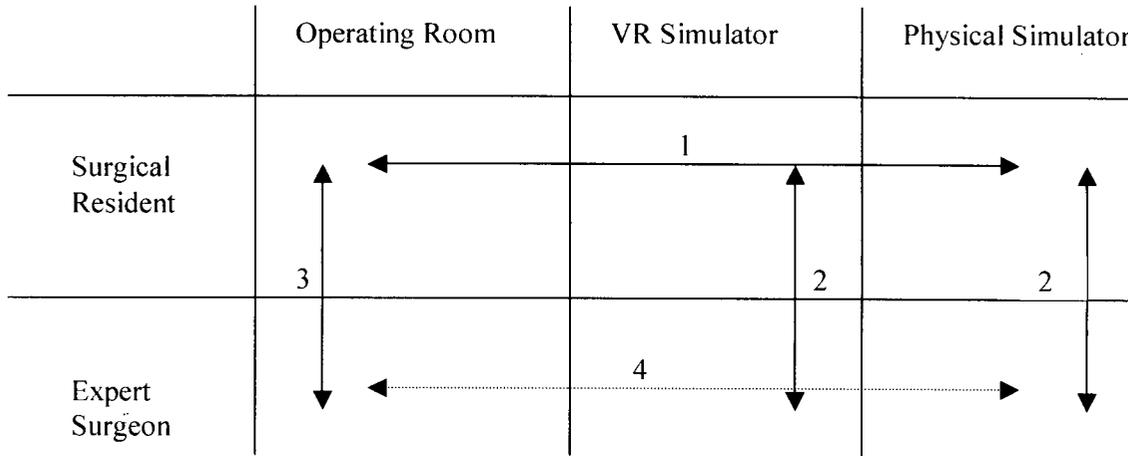
Experimental Methods for Assessing Validity of Laparoscopic Surgical Simulators

3.1 Introduction and Objectives

Historically, surgical education has been based on an apprenticeship style of training, where a senior surgeon would mentor the novice surgeons. It is now widely accepted in the surgical education field that this method of teaching is no longer acceptable (Feldman 2004, Rosser 1998, Winckel 1994). Inspired by the original flight training simulation programs that were used successfully in pilot training, surgical trainers and simulators were also created. The use of surgical simulators has come to the forefront of surgical education programs, and has been incorporated into some surgeon training programs (Fried 2004, Wentink 2003), allowing for unlimited unsupervised surgical practice. But from these studies, it has yet to be quantitatively shown how motor behaviour and patterns compare between the human operating room and surgical simulators; whether physically or virtual reality based. This study compares these motor behaviours in resident and expert surgeons using a custom-designed experimental system.

The primary objective is to study the performance, construct and concurrent validity of two types of surgical simulators: physical and virtual reality (VR), using surgical residents and experts as our subjects. Using our unique data collection and assessment system, we are able to quantitatively and objectively assess surgeon motor behaviours, and make comparisons to simulators using the same performance measures in all contexts. Our objective is to use the custom-designed surgical tool and data collection system described in the last chapter to collect motor behaviour data from novice surgeons in the human OR, and compare this to analogous tasks in surgical simulators to study performance validity of two surgical simulators. Also, if we can show that these simulators can distinguish between resident and expert surgeons, we can conclude that the simulators demonstrate construct validity.

For a diagrammatic representation of the goals in this project, please refer to Figure 3.1 below:



1 - Performance validity (residents)
2 - Construct validity (experts vs. residents)
3 - Concurrent validity: based on performance measures
4 - Performance validity (experts): completed by Kinnaird (2004)

Figure 3.1: Diagram of goals for this project. Each of the comparisons will demonstrate another type of validity. Goals 1-2 in bold are considered the main objectives. The dotted arrow line represents data from the study by Kinnaird (2004).

In the balance of this chapter, we describe the experimental methods used to study our objectives. We will describe the subjects, the settings, and the equipment used to collect and analyse the data. The methods of post-processing of the collected data are also covered, and our context comparisons to study performance, construct and concurrent validity.

3.2 Subjects and Settings

Three University of British Columbia PGY-4 (post-graduate year 4) surgical residents were assessed in three different settings (human operating room, virtual reality simulator, physical simulator) over a period of 5 months (March – July 2004). The residents consisted of 2 males, 1 female, all under the age of 35, and right-hand dominant. The surgical residents signed consent forms approved by the institutional review board to participate in our study.

Two expert surgeons' data that corresponded to the resident data was analyzed by Catherine Kinnaird (2004) in a recent study, and shared with this author. The expert surgeon data was collected by both this author and Kinnaird.

3.2.1 Settings

All of the subjects were evaluated in three settings: OR, VR simulator, and physical simulator.

3.2.1.1 Operating Room

The three surgical residents each performed a laparoscopic cholecystectomy under the direct supervision of an expert surgeon. Data was collected using the custom-designed data collection system with the experimental tool (as described in Chapter 2) during each of these procedures at the University of British Columbia Hospital between March and April 2004. The University of British Columbia (UBC) ethical review board gave approval for this data collection.

Equipment was sterilized as appropriate by ethylene oxide, and approved for use in the OR by the UBC Biomedical Engineering department. For the OR experiments, no prior selection of patients or staff was made. The patients were all required to have signed an informed consent form for the data collection prior to their procedure.

3.2.1.2 Virtual Reality Simulator

The virtual reality (VR) simulator used in our experiments consists of the Reachin™ Laparoscopic Training Package (Stockholm, Sweden) haptic feedback software, and the Immersion® (San Jose, CA, USA) hardware systems. The Immersion surgical station has two laparoscopic tools with interchangeable handles. These tools are similar to real laparoscopic tools in that they have four haptic degrees of freedom and a rotating tool tip. This hardware and software systems are complimentary and are combined to make our force feedback VR laparoscopic simulator. Generally, novices progress logically through increasingly difficult skill levels in training. The training package does not simulate an entire procedure, but the smaller tasks involved. There are tasks that are specific to the laparoscopic cholecystectomy procedure such as the camera placement, clip and cut, and dissection.

The specific module used in our experiments was the cystic duct dissection task of a laparoscopic cholecystectomy. The subject was to bimanually dissect away the surrounding fat

Two expert surgeons' data that corresponded to the resident data was analyzed by Catherine Kinnaird (2004) in a recent study, and shared with this author. The expert surgeon data was collected by both this author and Kinnaird.

3.2.1 Settings

All of the subjects were evaluated in three settings: OR, VR simulator, and physical simulator.

3.2.1.1 Operating Room

The three surgical residents each performed a laparoscopic cholecystectomy under the direct supervision of an expert surgeon. Data was collected using the custom-designed data collection system with the experimental tool (as described in Chapter 2) during each of these procedures at the University of British Columbia Hospital between March and April 2004. The University of British Columbia (UBC) ethical review board gave approval for this data collection.

Equipment was sterilized as appropriate by ethylene oxide, and approved for use in the OR by the UBC Biomedical Engineering department. For the OR experiments, no prior selection of patients or staff was made. The patients were all required to have signed an informed consent form for the data collection prior to their procedure.

3.2.1.2 Virtual Reality Simulator

The virtual reality (VR) simulator used in our experiments consists of the Reachin™ Laparoscopic Training Package (Stockholm, Sweden) haptic feedback software, and the Immersion® (San Jose, CA, USA) hardware systems. The Immersion surgical station has two laparoscopic tools with interchangeable handles. These tools are similar to real laparoscopic tools in that they have four haptic degrees of freedom and a rotating tool tip. This hardware and software systems are complimentary and are combined to make our force feedback VR laparoscopic simulator. Generally, novices progress logically through increasingly difficult skill levels in training. The training package does not simulate an entire procedure, but the smaller tasks involved. There are tasks that are specific to the laparoscopic cholecystectomy procedure such as the camera placement, clip and cut, and dissection.

The specific module used in our experiments was the cystic duct dissection task of a laparoscopic cholecystectomy. The subject was to bimanually dissect away the surrounding fat

and tissue from the cystic duct to expose it fully. The Maryland dissector is used in the right hand, and the surgeon's choice of grasper in the left hand. This is the typical surgical tool arrangement in the operating room.

3.2.1.3 Physical Simulator

The physical simulator used was a newly developed mandarin orange dissection. From literature searches, it was not found that any other laparoscopic surgical simulators use a mandarin orange simulator. Existing commercial physical simulators were not chosen for use as none readily represented the analogous dissections that were found in the OR or the VR simulator. The orange was chosen in consultation with expert surgeons, and met our requirements of a non-meat material as the same tool had to be used in the human OR (safety requirements do not allow surgical tools that are used on any other animal to be used in the human OR). The removing of segments of orange also represented the OR dissection most closely. The surgeons believed that this simulation was similar enough to real dissection tasks in terms of required movements and forces. The data from this simulator was collected by our custom-designed system as described in Chapter 2.

The subject used the instrumented Maryland dissector in their right hand, and was free to choose a standard tool for their left hand. The left hand tool was usually some type of grasper as similarly used in the OR. Generally, for right-handed surgeons, the useful tool is in the right hand, while the left hand is used more often for grasping and holding. The laparoscopic camera handler in these experiments was one of the researchers, with the subject directing to which direction to move and view. Using standard laparoscopic set-up (laparoscopic tower and camera) in a standard box trainer, the subject was asked to remove the peel and dissect out several segments of the orange using the experimental tool. They were specifically told to be cautious and to do as little damage as possible to the surrounding orange segments. As an indicator of face validity of this task, the head of surgical training at the University of British Columbia has decided to include this mandarin orange physical simulation in the surgical education program.

3.3 Performance Measures

The fundamental data available from our various systems include position and force data. The optoelectronic and electromagnetic systems provide us with 3D position and orientation data of the surgical tool. The force sensor and strain gauges give us force information. From this collected data, we then extracted a set of kinematics and force measures as described in the following sections.

The performance measures that are available are similar between the data from our collection system for the OR and the physical simulator as compared to the VR simulator (Table 3.1). Our data collection system was designed to allow for a broad range of measures to be taken. The VR simulator has built-in software that also gives many measures, but has limitations in the roll and tool tip force data as mentioned previously. We selected a variety of performance measures, and in the end decided to study a total of 26 measures to get a thorough understanding of the surgeon's motor behaviour. (The VR simulator gives us a total of 17 performance measures). The motions are all described in a reference frame at the surgical tool tip as seen in Figure 3.2.

3.3.1 Kinematics

3D kinematics data from the simulators and the OR are a performance measure studied in this project. The position data was differentiated to generate velocity, acceleration, and jerk data. This data can then be used to make comparisons between the three settings (OR, VR simulator, physical simulator). Specifically, the following kinematics performance measures were analyzed: velocity, acceleration, and jerk, in the axial, grasp, translate, transverse, absolute and roll tool tip directions. The VR simulator is limited in the tool tip roll direction and the force data as was mentioned previously.

3.3.2 Forces

Force data from the simulators and the OR were compared using the post-processed OR force data and the obtained simulators force data. Individual force components (axial, grasp, translate) and the transverse and absolute planes were analyzed. These components were calculated for the OR and physical simulator and available directly from the VR simulator

software for the VR data. Force data in the direction around the surgical tool axis (roll) were analyzed for the physical simulator and OR datasets, but were not available from the VR simulator software.

Table 3.1: Performance measures available from the three contexts. All measures are available in the physical simulator and OR contexts as data was collected with the experimental tool. The VR simulator is limited in roll and tool tip forces. Future software upgrades will allow for these measurements. See Figure 3.2 for tool dip directions. (Modified from Kinnaird 2004)

	Operating Room	VR Simulator	Physical Simulator
Tip Distance from Mean:			
Absolute	X	X	X
Roll	X		X
Tip Velocity: x,y,z	X	X	X
Transverse	X	X	X
Absolute	X	X	X
Roll	X		X
Tip Acc'n: x,y,z	X	X	X
Transverse	X	X	X
Absolute	X	X	X
Roll	X		X
Tip Jerk: x,y,z	X	X	X
Transverse	X	X	X
Absolute	X	X	X
Roll	X		X
Tip Force: x,y,z	X		X
Transverse	X		X
Absolute	X	X	X
Roll	X		X

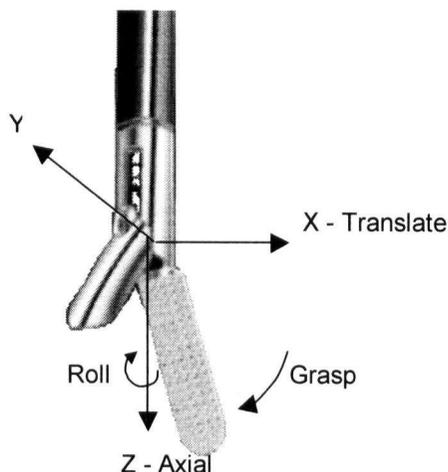


Figure 3.2: *Tool tip reference frame. The performance measures are taken relative to the surgical tool tip reference frame.*

3.4 Equipment Used

Each of the components of the sensor equipment used to collect the data was previously described in Chapter 2 section 2.3. To recap, for the OR and the physical simulator, we used an instrumented tool which incorporated optoelectronic and electromagnetic sensors to track position, and the force data was collected by a force sensor and strain gauge system. For the VR simulator, we used the data collected by the Reachin system. The remaining systems and the integration are described in the following sections.

3.4.1 Video Data

In addition to the sensors in the operating room, we also recorded videos of the surgery. Both an internal abdomen laparoscope camera view, and an external video camcorder focused on the surgeon were recorded. The two videos were time stamped and recorded onto standard VHS tapes using video-editing equipment. From these time stamped videos, correlations in time can be made with the collected kinematics and F/T data for analysis, and enabled us to identify the start and end points of the targeted tasks. The laparoscope video aided in the segmenting of the data, and start-stop points could be picked out for data segmentation as will be discussed in section 3.8.1.2. The external camcorder video allowed us to synchronize the data streams and for determining the characteristic synchronization movements needed as will be described in section 3.6.3.

3.4.2 System Component Integration

There are many components to this system that needed to be integrated to create a user-friendly system. This included all the sensors (F/T, optical, magnetic, strain gauge), and the video (laparoscopic and camcorder) and the computer to run these sensors (Figure 3.3).

All the systems except for the electromagnetic tracking system, Fastrak, were connected to a standard desktop computer (2.4GHz AMD Duron processor) with custom-designed data acquisition software written in Matlab (The Mathworks, Massachusetts, USA). Because of difficulties using multiple serial ports in one Matlab program, the Fastrak was connected to a separate laptop computer (minimum 800MHz) and the FTGUI software supplied with the Fastrak tracking system is used to collect data.

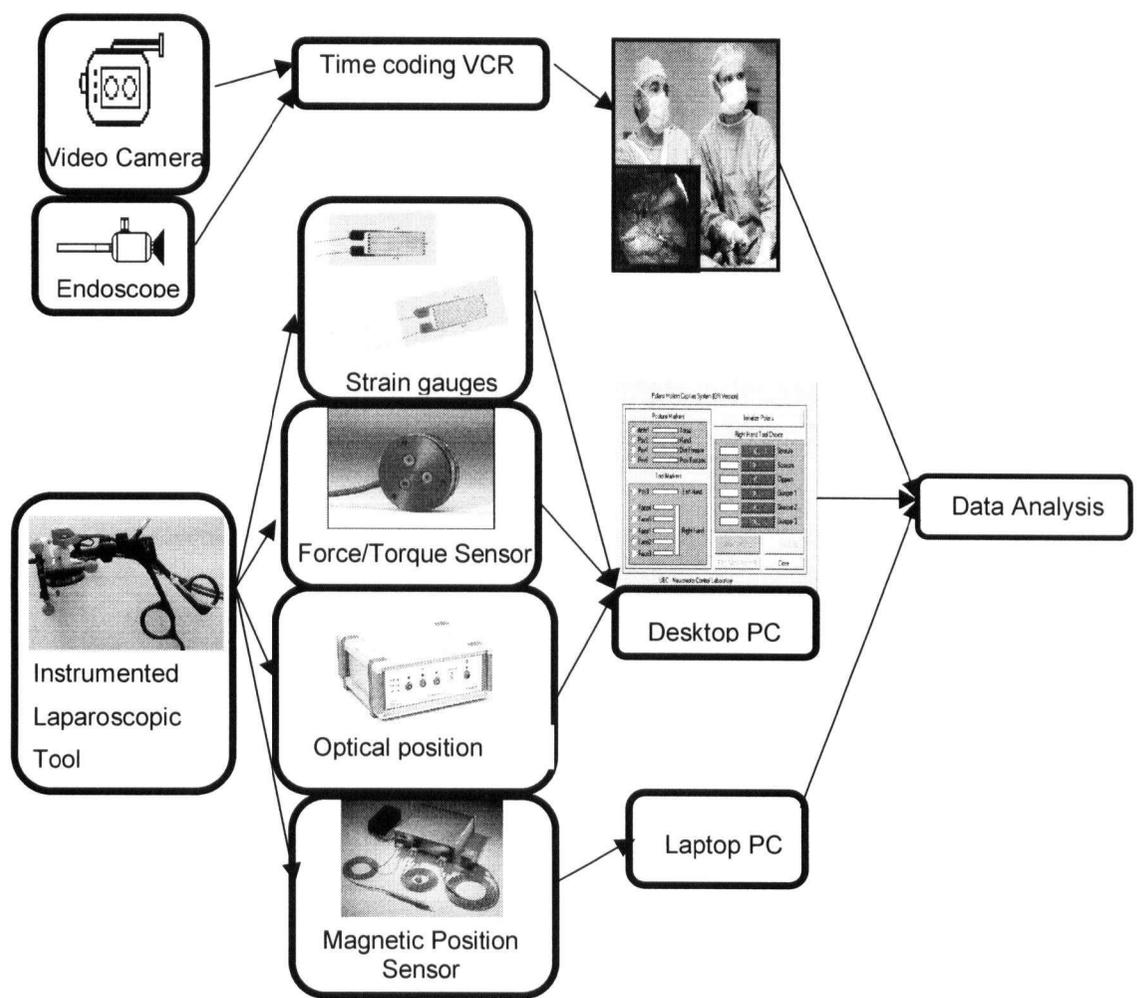


Figure 3.3: Components of the performance measurement system.

The standard laparoscopic surgical equipment was used in each OR procedure. The laparoscope system consisted of a standard 10mm – 0° surgical laparoscopic, camera and illuminator (Stryker Endoscopy). All equipment used for this study was approved by the Biomedical Engineering Department at the University of British Columbia Hospital, and was sterilized where appropriate with ethylene oxide.

3.4.3 Data Acquisition Software

Because of the variety of sensors used for data collection, various types of software were needed. Matlab was used as the primary data collection software because of its availability and usefulness in data collection and analysis.

The optical data was collected at 30Hz via a RS-232 serial port interface using existing custom-designed software implemented by McBeth in a previous study (2002). The graphical user interface (GUI) allowed for the user to see when the optical markers were visible or occluded, which allowed for better placement of the optical camera prior to the OR data collection.

Magnetic data was collected using the company (Polhemus) supplied data collection software (FTGUI) on a laptop computer. This data was collected at 120Hz through a RS-232 serial port interface.

The analog signal data from the strain gauges was gathered and converted to a digital signal using a Measurement Computing PCI data acquisition board. This board is supported by the Matlab Data Acquisition Toolbox and allowed for streaming strain gauge data at 120Hz.

Custom-designed Windows operating system drivers and Matlab functions previously created by Willem Atsma, a PhD student in our lab, were used to collect the ATI force/torque sensor data from the ISA data acquisition board that comes with the F/T sensor (Atsma 2001).

Streaming forces and torques could be collected at 120Hz.

Modifications were made to the original optical tracking software by McBeth to allow for data collection of the optical, F/T, and strain gauges all within the same GUI (Figure 3.4).

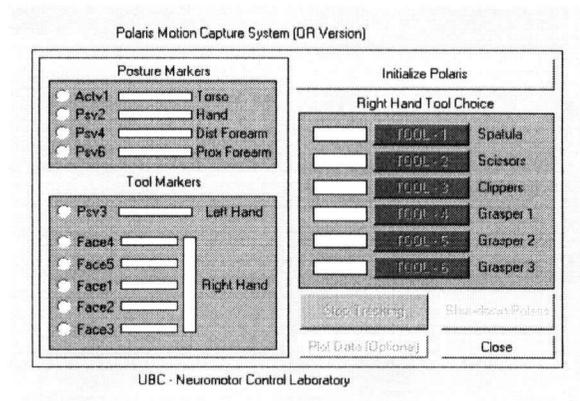


Figure 3.4: Custom designed data acquisition software. Polaris (optical tracking) GUI (original version by McBeth 2002) gathers streaming optical data, strain gauges and force/torque sensor with one button.

3.5 Data Collection

Study data was collected in the operating room from both virtual reality and physical simulators.

3.5.1 Operating Room Study

Each of the surgical residents performed a laparoscopic cholecystectomy at the University of British Columbia Hospital with an expert surgeon supervising. There were two researchers in the operating room (OR) for each experiment. One researcher scrubbed into the surgery to prepare the modified laparoscopic tool for use *in vivo*. This required cutting out and attaching a small thin section of OpSite™ surgical dressing to be used as a liquid barrier on the force/torque sensor. It was important to seal all crevices in the sensor to not allow any moisture to seep in. Also, a small piece of Mepore™ was used to wrap around the surgical tool handle where the strain gauges were mounted. This prevented the surgeon's fingers from getting caught on any edges or the strain gauge wiring, and kept the area clean and free from any foreign substances. The second researcher would help with set-up of the video camera and Polaris optical camera system, and then operate the computers and required software. The scrubbed-in researcher would also pass off the sensor wires from the surgical tool to the other researcher to be connected to the various computers and systems. If at any time the surgeon felt uncomfortable using the modified surgical tool, they could switch to a traditional non-modified

surgical tool. Immediately postoperatively, there were calibration “poses” needed with the modified tool. This calibration was used to synchronize and register the various streams of data. Also, a gravity vector was established with the tool in a neutral horizontal position to allow us to remove the gravity effects from the raw F/T data. The scrubbed-in researcher would hold and manipulate the tool in the required positions as data was collected. A more detailed explanation of the operating room protocol can be seen in Appendix A.

3.5.2 Simulator Data Collection

Both the surgical simulators were located in the Center of Excellence for Surgical Education and Innovation (CESEI) in Vancouver General Hospital (VGH). Each surgical resident came on separate days and completed the data collection on one of the two simulators on each visit. The VR simulator data was collected first, followed by the physical simulator at the later date.

The VR simulator data was collected three times in one session, and took approximately 20 minutes for all three trials. The physical simulator data was collected once, and took approximately 15 minutes to complete. The reason why there we were limited in the number of trials for each simulator was that the surgical residents did not have any more time to come in. The resident was asked to stand in a natural and comfortable position centred in front of the simulator, and to treat the simulation as an operative procedure. Before the start of each simulator data collection session, the surgeon was allowed a short familiarization and training session (~10minutes) on each of the simulators. This allowed the surgeon to be comfortable with the individual simulators and with the goals of the task, but did not allow for extensive practice or training.

Each resident completed the required task as we collected kinematics and force data with either our system (physical) or built-in software (VR). Post processing of the raw VR data was done with software designed by Iman Brouwer (2004), and produced continuous streams of kinematics and force data. This formatted VR data was similar to that gathered with our intraoperative system and allows for similar performance measure extraction, except that the VR data was limited in the force measures and roll in the tool tip direction. As was mentioned earlier, the x, y, and z direction force measures are available in the simulator defined “world” frame coordinates, but due to software complications, the proper transformation matrix was not

saved, and we could not transform the data to our tool tip reference frame. Therefore, we only could use the absolute force measure, and not the components. Roll torque is not available in the VR simulator.

3.6 Data Post-Processing

After data was collected in the operating room with the experimental surgical tool, many steps had to be taken to format the raw data into a usable form. See Figure 3.5 for a diagram of the post-processing steps.

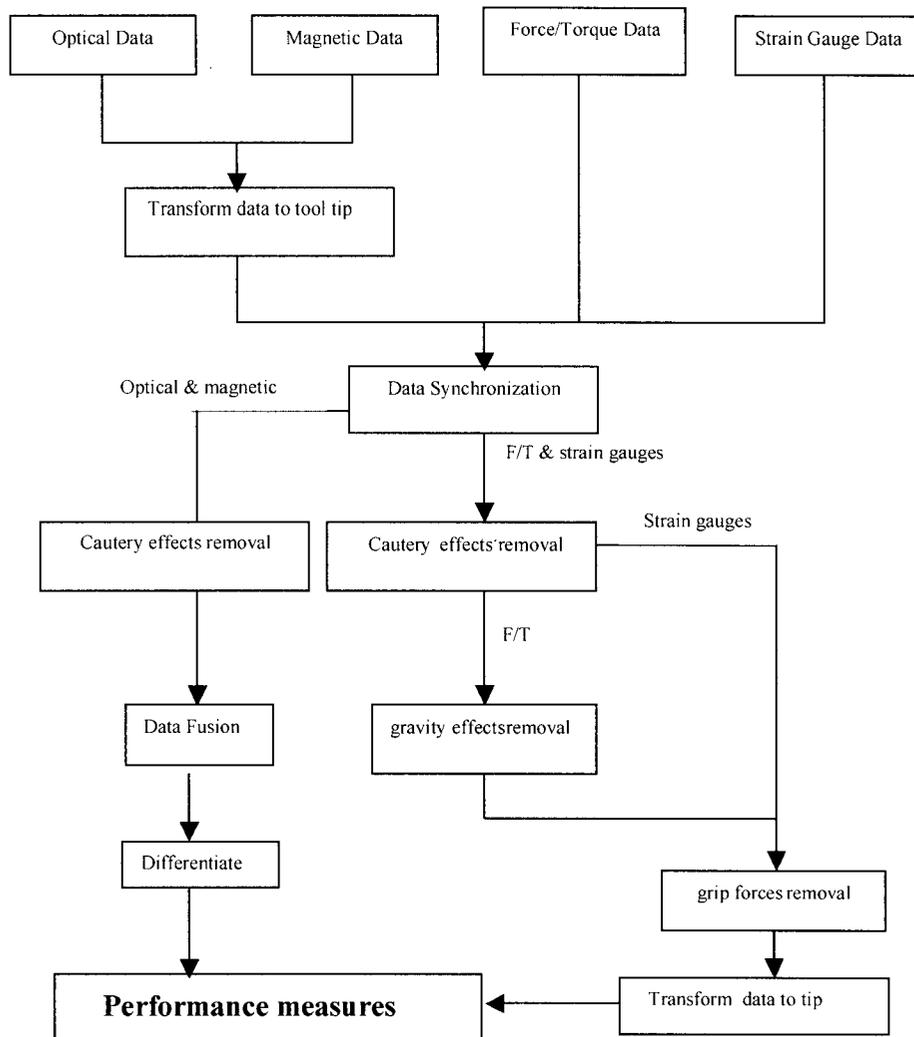


Figure 3.5: Data post-processing

3.6.1 Kinematics Data Registration and Calibration

We need to be able to represent the positional data gathered from both the optical and magnetic sensors in the same location in 3D space. Specifically, we want to know where the surgical tool tip is with respect to a world frame. As was described in Chapter 2 section 2.3.1, each of the position sensors tracks a 3D position using its' own tracking method. The Polaris optical tracking system can track five geometrically unique faces, which are set as three passively reflecting marker spheres that are custom-mounted on the array halo. Each of these faces represents a 3D frame in space, and the tracking camera will track one of their locations in space at a time, with respect to the camera reference frame. The Polhemus magnetic tracking system receiver also has its' own representation as a reference frame, and its location is tracked with respect to the transmitter reference frame.

As discussed earlier, we would like to fuse the two data streams from the optical and magnetic sensors. But to be able to do this properly, they need to represent the same locations in space. When using the experimental surgical tool, we are tracking the location of the surgical tool tip frame with respect to an anatomical body frame. A thorough discussion of the data registration between the optical camera and magnetic transmitter reference frames can be found in the thesis of Catherine Kinnaird (2004). Detailed information on component and OR system calibration and reference frame registration were also addressed in the mentioned work.

3.6.2 Force/Torque Data Registration and Calibration

This section will briefly outline the registration procedures for the 3D force/torque data. This is necessary to be able to produce force and torque measurements referenced to the surgical tool tip. The strain gauge data is used for estimating grip force and removing it from the tip force estimates and was previously described in Chapter 2 section 2.4.

3.6.2.1 Force/Torque Data Registration

After the raw F/T data was adjusted to account for gravity effects and grip forces, the F/T data was transformed to the surgical tool tip reference frame. The tool tip frame was established using optical and magnetic point probes and a calibration rig. The tip frame created here was established to be the same as in the kinematics registration. Further details of this can be found in Catherine Kinnaird's thesis (Kinnaird 2004).

3.6.3 Raw Data Synchronization

Because of the variety of sensors and computers used in this data collection procedure, the data streams were not initially synchronized and steps had to be taken to ensure that we could start the data streams at the same time to extract time-matched data. We therefore designed algorithms to allow us to synch the various sensors (Figure 3.6). As described in the Operating Room protocol (Appendix A), a large characteristic movement was made by the surgeon at the end of the surgery, which enabled us to find corresponding times in the position datasets. This characteristic move was much larger than anything that would be seen during typical surgical movements. Also as part of the protocol, the surgical tool was held in a horizontal stationary position before and after the characteristic large move to further differentiate this synching movement from the surgeon's regular tool movements. The now synched position data was synched to the ATI force data by a "hit" against a surface. Lastly, the synched position and force data was synched to the strain gauge data by a large "squeeze" to the tool handles.

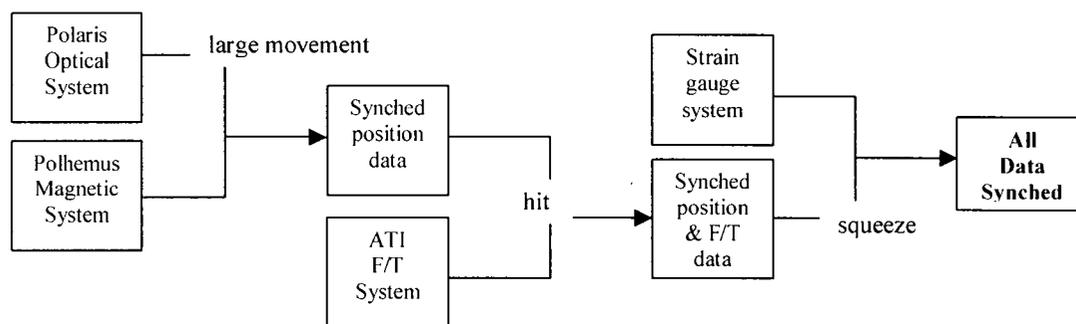


Figure 3.6: *Data synchronization process. The position sensors are synched first by visual inspection of data for the large characteristic move. This is then synched to the F/T data by looking for the large "hit". This synched data is then time synched with the strain gauge data by the large squeeze that is seen in the strain and force data, and all data is now synchronized in time.*

To actually synch the optical and magnetic kinematics data, a visual inspection of the position data during the large characteristic move is done. A small segment of time is chosen from both positional datasets, and an optimization routine is executed. From this small window in time, an initial guess of Δt is made and input into the algorithm. The real Δt is calculated by using a non-linear least squares optimization routine found in the Matlab Optimization Toolbox.

Further details on this can be found in Catherine Kinnaird's thesis (2004).

The synched kinematics data then is synched to the F/T data by again visual inspection and a window of time is chosen. The large characteristic move also includes a "hit" as described earlier, which is recorded by the sensors. This "hit" is larger than any typical forces in surgery. The F/T data can then be synched with the kinematics data. Finally the strain data was synched to the previously synched kinematics and F/T data. The characteristic move includes a large squeeze of the tool handles, which is after the big "hit" and usually larger than any squeezes that a surgeon would do.

The external camcorder video and the internal laparoscopic video also needed to be time synched with the collected data. These two videos are collected separately and could be synchronized as we could see the surgeon inserting the laparoscopic camera into the trocar in both videos. Our external camcorder was focussed on the main surgeon, and all their movements were recorded. The internal laparoscopic camera also recorded continuously. Once this insertion movement was identified, then video-editing equipment could be used to time-stamp both internal laparoscopic video and external camcorder video.

These videos then had to be synchronized with the collected sensor data. By visually inspecting the magnetic positional data, we could see when the surgeon removed the experimental tool and laid it down. For example, the surgeon removes the experimental Maryland dissector to use the clipping tool. These times could be seen both in the external camcorder video and in the magnetic positional data stream, and this information could be used to synchronize them together.

3.7 Electrosurgery Unit

An electrosurgery unit (ESU) is a common and typical piece of equipment in today's modern OR. It allows the surgeon to cut through tissues while coagulating any blood vessels at the same time. This is beneficial for both the patient and the surgeon. The patient will lose less blood when cuts are made this way, and the surgeon is able to operate in an almost blood-free environment.

The ESU delivers radio frequency (RF) currents, which allow the surgeon to cut, cauterize or coagulate live human tissues. An electrical wire from the ESU is attached to the surgical tool through the port. The electrical current passes down through this connection, down the innermost shaft of the surgical tool and through to the desired tissues via the surgical tool tip. The monopolar type of ESU was used in these experiments. The monopolar ESU requires that the electrical current pass from the active electrode through the body, and exit through a passive electrode attached pre-operatively to the patient's body.

There are number of settings that can be chosen on a typical ESU. There are generally two modes: cut and coagulate. Generally in these applications, the ESU cut mode is capable of a 400KHz (1200V) voltage. And in the coagulation mode, 250KHz (3500V) at 40KHz bursts is available for surgeon use to coagulate tissue. In the OR, a surgeon will usually request the "blend" setting that is a combination of cut and coagulation settings. This allows for cutting through the tissues while coagulating any blood vessels along the way.

3.7.1 ESU Effects

Because of the variety of instrumentation and sensors that we have attached to the experimental tool and because we had cut the original tool to add our modifications, we needed to ensure that our sensors would not be damaged and that the current would pass uninterrupted through the tool. Preliminary tests with a typical OR ESU borrowed from Vancouver General Hospital Biomedical Engineering were completed to see the effects on all the sensors.

The first tests were completed to determine if the use of cautery would damage the sensors. We were unsure if the electrical current was strong enough to permanently damage the sensors. We did incrementally increase the voltage and current output to maximum from the ESU, and recorded the data and checked the sensors' operation. We found that no sensors would be damaged, but the readings from the strain gauges, magnetic position system, and the F/T data would all be affected by a significant amount of noise while using coagulation and blend modes. The cut setting did not have any noticeable effect on the data.

3.7.1.1 Removal of ESU Effects

The magnetic, strain gauges and F/T data are all affected adversely by the ESU. Although, the amount and degree of noise is different for each sensor, the basic approach to dealing with and removing the noise and extracting the proper data is very similar for each sensor, with small modifications and adjustments made for each. But according to our experimental OR data, cautery may be applied for as long as 15 seconds at a time.

The effect of ESU activity on strain gauge data is shown in Figure 3.7.

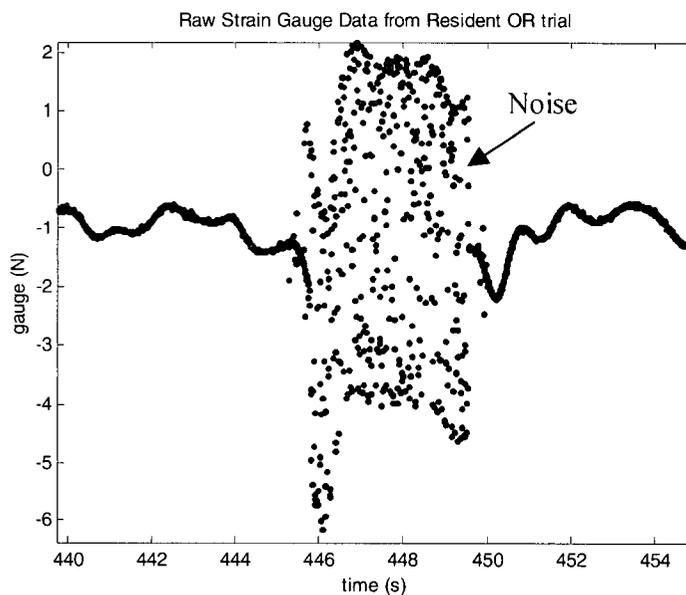


Figure 3.7: Strain gauge data with electrocautery noise.

It is obvious that when the ESU is applied, the strain gauge data is completely distorted, and we felt (after some experimentation) that no amount of filtering would produce a useful signal. We decided to simply remove these noisy sections from our data.

The data removal algorithm is based on looking at small increments of time ($\sim 1/10$ s) and comparing it to the small time segment before it. If the difference between these windows is beyond a given threshold (chosen by comparison of the known good data from noisy data), the noisy data is removed (Figure 3.8). The threshold values varied depending on the data, and were chosen by examining data surrounding the noisy section. If a large amount of data were

affected in a block, then the whole block would be manually removed. This ensured that minimal data would be removed. We were always careful during data removal, and generally under-removed data as opposed to over-removing it.

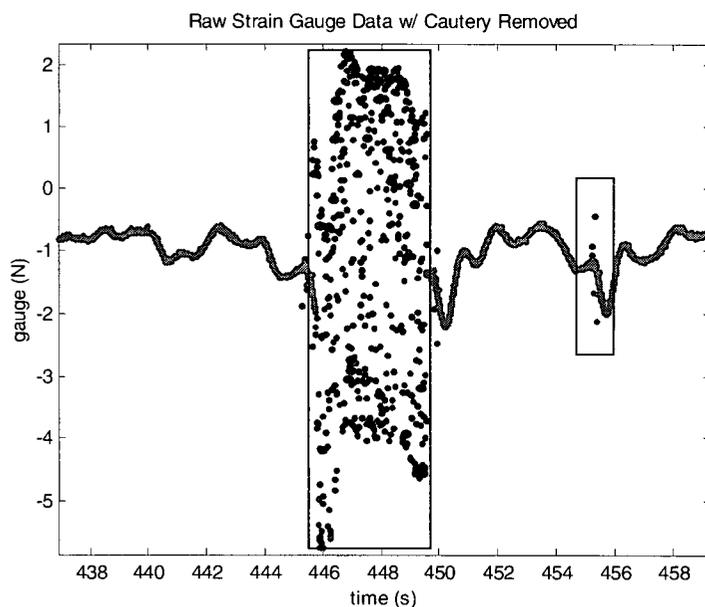


Figure 3.8: *Raw and noise removed strain gauge data. Large and small sections of noise are removed.*

The magnetic position tracker is also affected by the ESU (Figure 3.9), but a built-in feature of the Polhemus FTGUI software is its ability to track when errors occur, and make a record of these errors. This is seen in the raw data output. This allows for an easier preliminary data removal in the magnetic stream, as this erroneous data can be removed. Any remaining noise artefacts can be removed manually or by using the filtering technique of monitoring the sudden noisy changes as described above.

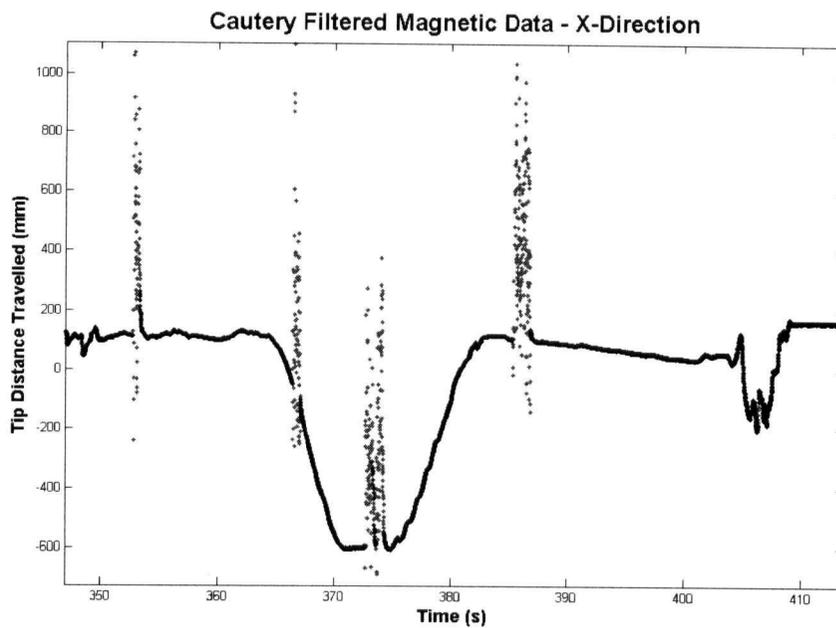


Figure 3.9: *Electrocautery affected magnetic data.*

Lastly, the F/T dataset was also affected by the use of ESU. As can be seen in Figure 3.10, it is quite obvious when the cautery current is used, as the data suddenly changes showing large spikes. Data removal was done manually or by using the sudden change in profile filtering method similar to that of the previous sensors.

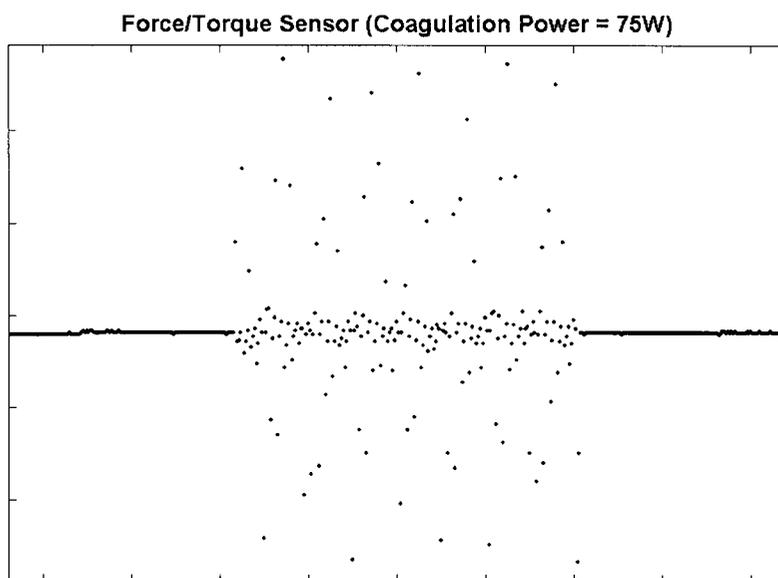


Figure 3.10: *Electrocautery affected F/T data.*

3.8 Task Comparisons

All surgical procedures are highly variable and we are in need of a method of making comparisons between these types of procedures. We also need a method to compare these OR measurements to both the measurements from the VR and physical simulators to be able to assess the validity of the simulators and our performance measures. In our OR studies, data was gathered from laparoscopic cholecystectomies (gallbladder removals) without prior selection for patient or operating room staff, therefore increasing the variability between the procedures. In contrast, the VR simulator provides a very structured, rigid and repetitive environment for teaching and evaluation of surgical skills. The tasks are broken down and each one can be practiced separately (i.e. clipping task, dissection task, etc.). The physical simulator is also a relatively repetitive and repeatable environment. But to compare between these three settings, we need to be able to extract similar data from each context.

3.8.1 The Dissection Stage

The experimental tool chosen in consultation with expert surgeons was the Maryland dissector (or grasper). This tool was selected as it is used extensively throughout the first portion of most laparoscopic cholecystectomies, and most surgeons are comfortable and familiar with its use. We have selected a commonly completed stage in the operative procedure to demonstrate our approach to performance evaluation. This is the dissection stage, and is a key component in the laparoscopic cholecystectomy procedure. The dissection stage of the laparoscopic

cholecystectomy involves removal of extraneous tissues and fat surrounding the gallbladder and the vessels (cystic artery and cystic duct), and to isolate these vessels for clipping and cutting.

Both the VR and physical simulators have analogous tasks that can be compared to this dissection stage in the actual human operation. The VR simulator has a cystic duct dissection simulation where the surgeon must dissect away the fat surrounding the gallbladder and cystic duct and artery (Figure 3.11). The physical simulation is the dissection of mandarin orange fruit using the hybrid experimental tool and data collection system.

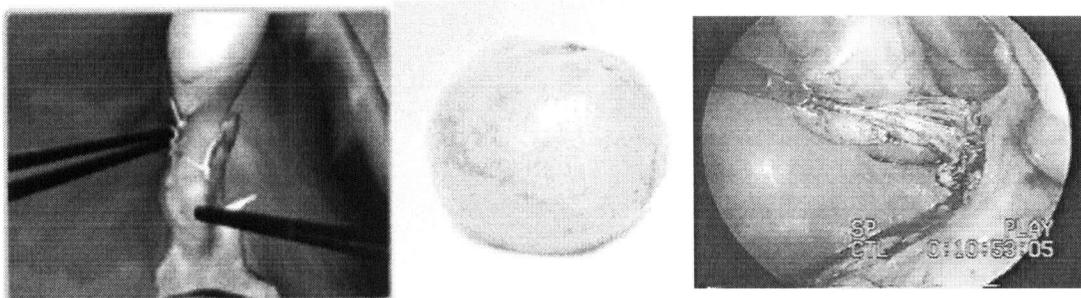


Figure 3.11: *VR simulator vs. Physical simulator vs. OR. From left to right: VR simulation of cystic duct dissection. Physical simulator is an orange. Actual OR dissection task.*

3.8.2 Data Segmentation

In a typical OR experiment, we would collect approximately 10-15 minutes of OR data from all our sensors. This would lead to very large raw datasets. In order to better manage these large amounts of data, and to be able to break down the procedure according to the hierarchical decomposition, data segmentation was used. The post-processed raw data could be segmented with the help of the time-stamped internal laparoscopic video and the external camcorder video. We were specifically concerned with the dissection task of the procedure. After data segmentation, performance measures could be extracted, and then compared with analogous tasks in the VR and physical simulators.

3.8.2.1 Data Segmenting

The internal laparoscopic synchronized and time-stamped OR video was used to create a start and end point for each dissection task. Each start point was taken as the moment in time when the experimental surgical tool first contacts the tissues. The end point was when the tool was

removed from the tissue, and taken out of the trocar. These start and end points are then used to segment out the dissection task of both the formatted kinematics and F/T data.

For each procedure, the dissection task of a typical laparoscopic cholecystectomy is of interest. This dissection task is decomposed further into segments to allow for easier data manipulation. These segments are identified by visual observation of the kinematics, F/T and video data to see when the surgical tool tip is in contact with the tissues and is being actively used. Generally, the first segment is when the surgeon has first entered the surgical tool into the abdomen is exploration and anatomy identification (i.e. cystic duct, cystic artery, surrounding vessels). The cystic artery and cystic duct are identified, separated, clipped and cut, respectively. Our experimental tool is used extensively for these portions of the procedure. Once the cystic duct is cut, the surgeon tends to use a hook or spatula tool to dissect and remove the gallbladder from the liver rather than the experimental tool.

The time and parts of each procedure where the surgeon uses the experimental tool is different between surgeons, and can vary between procedures. Usually, the most variable portions of the procedure are isolating Calot's triangle, and dissecting the gallbladder. This was usually due to a chronically inflamed gallbladder resulting from a patient waiting a long time before having the surgery. This sometimes led to longer operating times.

For the VR and physical simulators, no data segmenting was done, as the entire task was considered to be dissection. We did consider all three contexts to be analogous in that we were able to separate out the dissections tasks in the OR data, and the two simulators only included dissection task data.

3.9 Setting Comparisons

After data was collected and formatted from the three settings (OR, VR and physical simulator), comparisons could be made between these contexts. We wanted to examine intersubject, intrasubject and context differences and similarities. The novice surgeon data also needed to be compared to the expert surgeon data previously collected and analysed (Kinnaird 2004). The raw data consisted of time histories of displacement, velocity, jerk and force acquired over intervals of up to approximately 20 minutes. The main point being that a large

amount of data was collected from somewhat different unstructured (variability in performance) contexts, therefore we cannot make detailed specific comparisons. To assess differences, we chose to use a statistic that would be sensitive to any differences between the cumulative probability distributions (CPD's) of the performance measures. The Kolmogorov-Smirnov (KS) statistic was used in previous studies in our lab, and we have decided to continue with its use.

3.9.1 Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov (KS) statistic is a parameter free measure of the difference between two CPD's. It requires the data from two cumulative probability distributions (CPD's) of performance measures such as velocity, force, etc., and it is the maximum absolute vertical difference (D) value between the two CPD's (Figure 3.12). The D -value ranges from 0 (similar) to 1 (different). Another advantage of the KS statistic is that it makes no a priori assumptions about the shape of the CPD's (i.e., they do not have to be Gaussian) and is relatively insensitive to outliers in the data (Hodgson 2002). This characteristic is especially important as our data does include many outliers, even after filtering. For these reasons, the KS statistic has been found to be a valuable tool in evaluating behaviours in different environments (Boer 1996, McBeth 2001), so we have chosen to use the KS statistic for all our contextual comparisons.

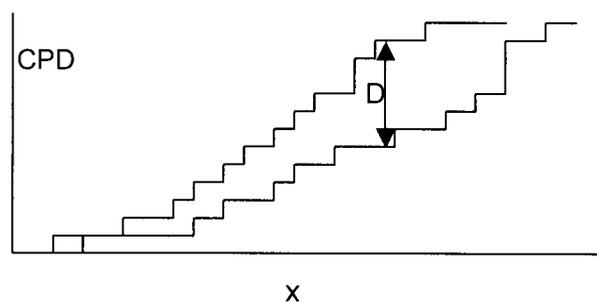


Figure 3.12: *Kolmogorov-Smirnov CPD. Comparison of cumulative probability distributions to find D -value of the KS statistic.*

3.9.2 Comparisons

In order to better understand the usefulness of the KS statistic in our application, a brief description of the comparisons done is needed. (A more thorough description of these comparisons can be found in Chapter 4 section 4.2.1).

We have collected data with 5 subjects in 3 settings (OR, VR and physical simulator). The 5 subjects are divided into 2 levels: resident and expert. We have made intrasubject, intersubject, interlevel, and intersetting comparisons. For example, if we wanted to compare one resident to another in the performance measure of force, we would create the CPD of force for each resident, and then be able to find the KS D-value. The D-value gives us the difference between these two subjects. The larger the D-value, the larger the difference between subjects.

3.9.3 Assigning Confidence Intervals

When we express a difference between two CPD's, we must also compute a corresponding confidence interval (CI) on the difference measure. Although difficult to do analytically, it is comparatively straightforward to compute using bootstrapping methods.

Bootstrapping is a computationally intensive method that involves using the sample actually obtained as an estimate of the underlying distribution, and randomly resampling the dataset and re-computing the KS statistic at each bootstrapping cycle, which will give us a measure of the accuracy of the D-value by assigning a confidence interval to it. Bootstrapping tries to recreate the relationship between "population" and "sample" by assuming that the sample available (i.e., OR and simulator data) is representative of the underlying population (Efron 1986). The bootstrap method estimates are used to give an estimate of the measurement error on the D-value calculated for that specific case.

3.9.4 Dependent Data and the Moving Block Bootstrap

Simple bootstrapping methods are based on the assumption that the data are completely independent from each other. In our case, however, the value each data point is highly correlated with its neighbours (x_{i-1} , x_{i-2} , ... x_{i-m} where m is unknown) because they come from a continuous stream of data. This temporal correlation implies that there are effectively fewer

independent data points, so applying the standard bootstrap method will result in unrealistically tight confidence intervals.

The general bootstrap method is more complex with time-correlated data, but the basic ideas remain. The moving block bootstrap (MBB) method was chosen as this technique accounts for dependent datasets (Kunsch 1991, Liu 1992). The MBB resamples blocks of consecutive data at one time, as opposed to resampling a single observation as is done in the standard technique. This results in the dependent structure of the original data block being retained within each resampled block.

The MBB is applied to our dependent data as shown in Figure 3.13. The original dataset $X_n = \{X_1, \dots, X_n\}$ is partitioned into overlapping blocks of length l to create a matrix of blocks $\{\beta_1, \dots, \beta_N\}_{N \times l}$. From this matrix of overlapping blocks, a suitable number of blocks k is resampled with replacement to make a resampled set of blocks $\{\beta_1^*, \dots, \beta_k^*\}$. The new data X^* is then assembled from the elements of β^* . The value of k is chosen so that each bootstrap sample is the same length as the original sample data ($n = k * l$). The size of each block of length l , increases with the length of the original sample. The value of l should be on the order of $n^{1/5}$ (Hall 1995).

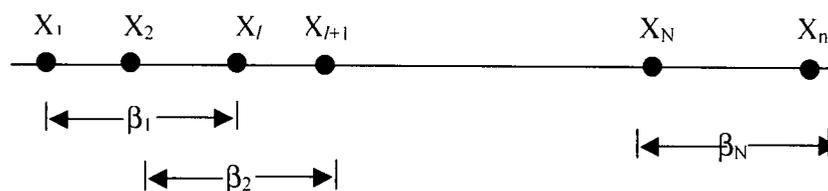


Figure 3.13: *Moving Block Bootstrap. The MBB method breaks the dependent dataset that is to be resampled into $N = n - l + 1$ overlapping blocks. These blocks are then randomly resampled with replacement to length k and the resampled dataset then assembled from the resampled blocks, thereby preserving the dependent structure of the original dataset.*

3.9.4.1 Measurement Resolution

We have calculated many D-values from our CPD's, and we need to know how reliable those D-values are. We use the MBB method as described above to assign a confidence interval to each of our D-values. Using the MBB we resample each CPD and this results in many (i.e., 1000) D-values, and we can create a CPD of D-values. We then can assign a confidence

interval on D_{1-2} from the 2.5-97.5th percentiles of this CPD of D-values (Figure 3.14). This confidence interval shows the range of D-values that are likely to occur with the underlying distribution. The size of the confidence interval is dependent on the effective size of the dataset and variability within the distribution. Even taking data dependency into consideration, we still do have large datasets, and small confidence intervals are expected for a D-value calculated between two distributions. The confidence intervals give us an estimate of the measurement error involved in calculating the D-value for each comparison.

An example would be to measure the circumference of one green apple and one red apple with an inaccurate tape measure, the confidence interval for this measurement would give us a value related to the measurement technique and we could not make any assumptions about the circumference's of all red and green apples. Therefore, we should keep this in mind when examining our calculated confidence intervals.

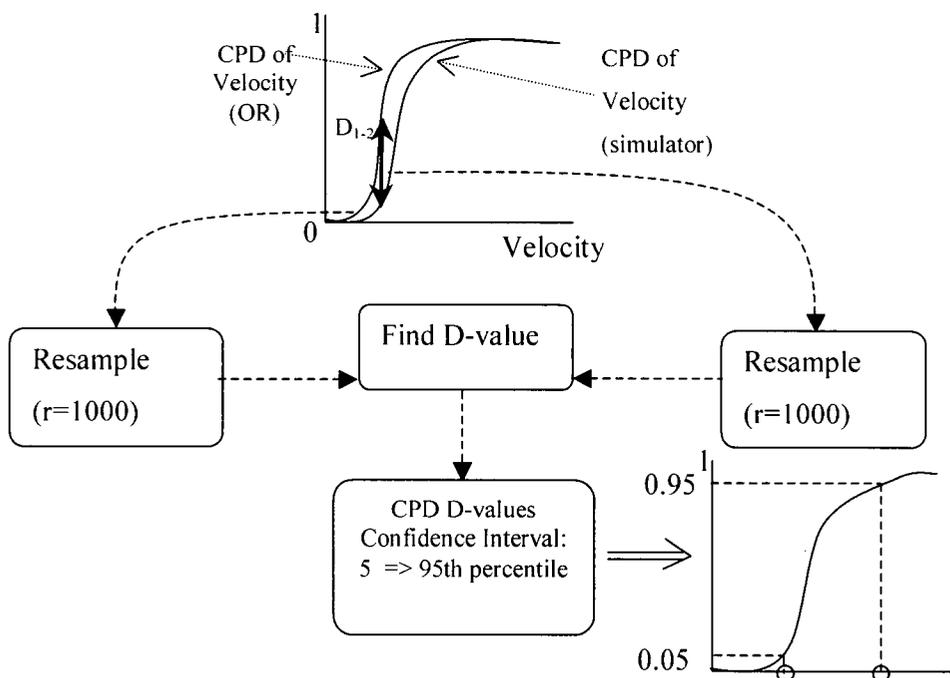


Figure 3.14: Confidence intervals for D-values. To assign a confidence interval to a D-value, each CPD is resampled to create a CPD of D-values (Source: Kinnaird 2004)

A measure of statistical significance is necessary to judge how much different our performance values are. One of the CPD's is assigned as the reference (CPD_{ref}). Each performance measure (i.e. velocity, force, etc) has its own CPD_{ref} . This CPD_{ref} is then resampled and each resampled distribution (CPD_{RS}) is compared to the CPD_{ref} . This is done many times (~ 1000) to get a distribution of D-values between the reference and resampled CPD (D_{RS-ref}). The D-value at the 95th percentile of the $CPD(D_{RS-ref})$ is the critical D-value (D_{cr}) (Figure 3.15). The D_{cr} is used to identify statistical difference between any measured D-value (D_{meas}) and the reference if D_{meas} is greater than D_{cr} . For example, if we are investigating the force profiles of two surgeons in the OR, surgeon 1 is considered the reference and is resampled to get $CPD(D_{RS-ref})$. If $D_{surgeon 1 - surgeon 2}$ is outside the 95th percentile of $CPD(D_{RS-ref})$ then, the two surgeons' force behaviours are different.

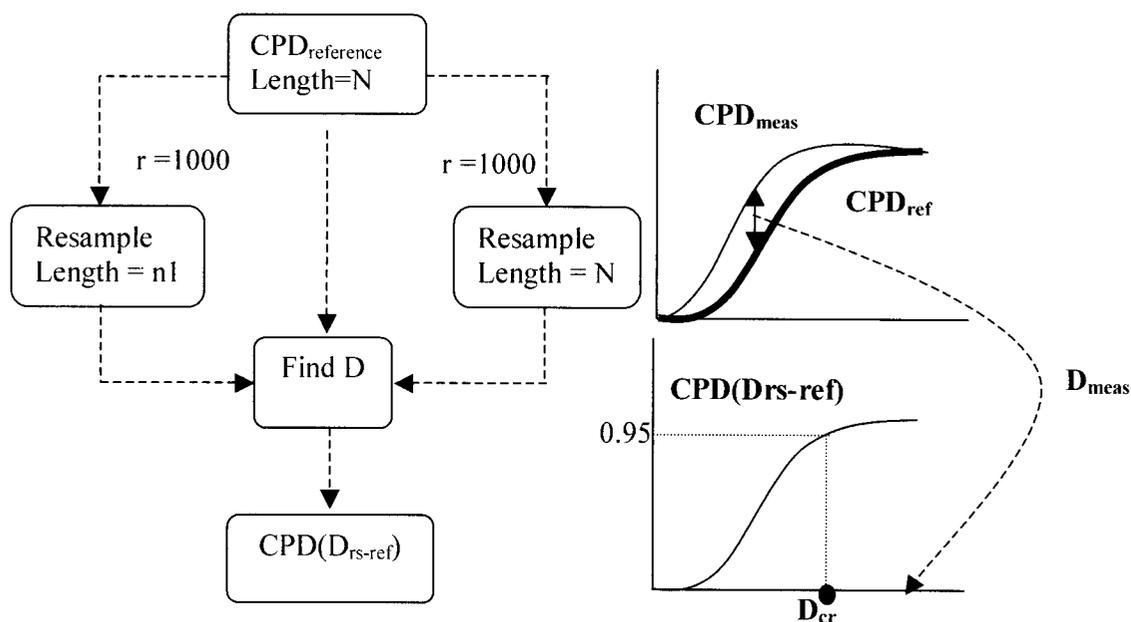


Figure 3.15: CPD of D-values. Finding a CPD of D-values between the CPD_{meas} and CPD_{ref} : we can assess the relevance of any measured D-value. If D_{meas} is greater than the D_{cr} value of $CPD(D_{rs-ref})$ then the two CPD's under consideration are different (Source: Kinnaird 2004).

3.10 Discussion

The Maryland dissector tool was chosen as the experimental tool in consultation with expert surgeons. This particular tip was selected as it is frequently and commonly used in

laparoscopic cholecystectomies. The Maryland tool tip is interchangeable, as future studies may require other useful tool tips.

A custom designed and built bracket was created to mount the various sensors required for this project. By creating this bracket and mounting it to the surgical tool shaft, all sensors were mounted securely and kinematics and F/T measures could be extracted for study.

The use of the electrosurgical unit (ESU) in the OR caused a significant amount of distortion and noise in our raw data. The magnetic, F/T and strain gauge sensors were all adversely affected by the ESU. This effect required data removal to be done before performance measures could be analyzed. A technique to monitor the velocity changes was used to successfully remove the affected sections of noisy data. The remaining ESU affected data could then be removed manually.

Forces and torques of the surgical tool tip were collected using the experimental set up and a tri-axial transducer mounted on the bracket. Many hours of calibrations and data registration were done in post-processing to remove gravity effects, grip effects, and electrosurgery unit effects.

Chapter 4

Results of a Quantitative Study to Assess Laparoscopic Surgical Simulator Validity

4.1 Introduction

In this chapter, we present pilot study data illustrating intersubject variability, intrasubject differences, and the reliability of our chosen performance measures. The performance and behaviours of the novice surgeons were compared to each other in OR and simulators (i.e., performance validity), and then again compared to the experts (i.e., construct validity). We also analyze the concurrent validity of the simulators based on our performance measures in the OR as the gold standard. The implications of the analysis are then discussed as concerning the reliability of our data collection system and construct validity and performance validity of the simulators.

The protocol as outlined in Chapter 3 section 3.5, and the lengthy post-processing (Chapter 3 section 3.6) were followed for each of the three OR procedures, and physical simulator data collections. This resulted in time synchronized and post-processed kinematics and force data referenced to a common reference frame at the surgical tool tip. The dissection task data of the surgical procedure is also broken down into segments as discussed in Chapter 3 section 3.8.1.2.

4.2 Results

We were able to successfully collect OR data 3 times (one surgery from each resident). We also collected data in the virtual reality (VR) and physical simulators. A summary of the data collections is shown below in Table 4.1

Table 4.1: *Summary of successful data collection from each context*

	OR	VR Simulator	Physical Simulator
Resident 1	1	3	1
Resident 2	1	3	1
Resident 3	1	3	1

4.2.1 Context Comparisons

The comparisons between the contexts and the subjects are completed in many different areas. We have collected data from the surgical residents in each context (OR, physical simulator, VR simulator), and will make comparisons within these. Then this data will be compared against the expert data previously presented by Kinnaird (2004).

4.2.1.1 Surgical Residents

The first comparisons (Figure 4.1) presented will be intrasubject from each procedure. Each procedure is divided into segments as discussed in Chapter3 section 3.3.1.2, and these segments compared to each other ($\Delta 1$). This will investigate intrasubject intraprocedure variability and repeatability. Next, the intrasubject intertrial VR ($\Delta 2$) comparisons are shown to investigate repeatability in the VR simulator of the residents. Thirdly, the intersubject intrasetting ($\Delta 3, \Delta 4, \Delta 5$) results will be analyzed. Each of the residents will be compared in the three settings (OR, physical simulator, VR simulator) to evaluate consistency at the skill level. And lastly, the intrasubject intersetting ($\Delta 6$) results will compare each of the residents' behaviour in the OR to the simulators to performance validity of the two simulators.

	OR	Physical Simulator	VR Simulator
Resident 1	① ←————→ ⑥		② —————→
Resident 2	① ←————→ ⑥		② —————→ ⑤
Resident 3	① ←————→ ⑥		② —————→

$\Delta 1$: Intrasubject intraprocedural OR
$\Delta 2$: Intrasubject intertrial VR simulator
$\Delta 3$: Intersubject intrasetting OR
$\Delta 4$: Intersubject intrasetting physical simulator
$\Delta 5$: Intersubject intrasetting VR simulator
$\Delta 6$: Intrasubject intersetting (OR versus VR, OR versus physical)

Figure 4.1: Context comparisons for surgical residents. The numeric values in the table represent the respective Δ s.

4.2.1.2 Expert Surgeons

To study construct validity as stated in our objectives, we need to compare our surgical resident data to that of the expert surgeons. The expert surgeon data was collected and performance measures extracted by Catherine Kinnaird (2004). Our resident to expert comparisons will be called “interlevel” comparisons (Figure 4.2).

This comparison will be interlevel intrasetting. This will demonstrate the results of the experts compared to the residents in each of the contexts (OR, VR simulator, physical simulator). $\Delta 7$ is new method of evaluating concurrent validity as we have OR expert data as the “gold standard”. We have the same performance measures available in each of the other contexts (resident skill level and simulators). This way we are able to quantitatively make suggestions in the concurrent validity of the simulators. $\Delta 8$ and $\Delta 9$ allow us to investigate the construct validity of both the physical and VR simulator, as we are trying to detect skill level differences.

	OR	Physical Simulator	VR Simulator
Residents	↑ ⑦	↑ ⑧	↑ ⑨
Experts	↓	↓	↓

$\Delta 7$: Interlevel Intrasetting OR

$\Delta 8$: Interlevel Intrasetting physical simulator

$\Delta 9$: Interlevel Intrasetting VR simulator

Figure 4.2: *Interlevel context comparisons for experts and residents.*

4.2.2 The D-Value

The KS statistic D-value is calculated for all context comparisons. The D-value depends on the size of the original sample sizes of the distributions. Our sample sizes are all in the magnitude of several thousand data points, and the larger the sample size the smaller the D-value must be to be considered “similar”. Generally, when a dataset is resampled from itself (D_{rs-ref}), D-values are usually about 0.02-0.05. (Remember that a D-value of 0 is similar, and a D-value of 1 is maximum difference) When two CPD’s are different, we usually see values of 0.8-1.

4.2.3 Presentation of Results

The performance measures as discussed earlier are velocity, acceleration, jerk and force in the six tool tip directions: axial (z), grasp (y), translation (x), transverse ($\sqrt{x^2 + y^2}$), absolute ($\sqrt{x^2 + y^2 + z^2}$), and roll about the tool axis. The performance measure of distance from the mean (D mean) is presented only in the absolute and roll about the tool axis directions as it is sensitive to the choice of location of the global reference frame.

The cumulative probability distributions (CPDs) of all twenty-six performance measures in all directions are presented in a large plot with twenty six subplots. The 75th percentile of the data is shown for better visualization of the results as this area shows the critical areas of the CPD's. The important differences between CPD's are always in this region. The CPD's all have long tails, and if the entire CPD was shown, the critical areas would appear to be vertical, and the important differences would not be easily seen.

The D-values of the comparisons is also calculated and presented in another plot. The D-values are shown with confidence intervals, and the critical confidence interval is also shown for finding the statistical difference (CPD(D_{RS-ref})).

4.2.3.1 $\Delta 1$: Intrasubject Intraprocedural OR Comparisons

Each of the surgical residents performed a laparoscopic cholecystectomy with an expert surgeon in attendance and supervising. Each resident had one session of data collection in the OR, and the results from each are presented in the following sections.

Each surgical dissection task was divided into three segments to examine intraprocedural repeatability. The first segment consisted of anatomy exploration and identification. The second segment was the cystic duct and artery dissection. And the third segment was the gallbladder removal from the liver bed. We investigate the repeatability of the resident within one procedure.

4.2.3.1.1 Resident 1: Intrasubject Intraprocedural OR

The performance measures for the three segments of the OR procedure are extracted (Figure 4.3). In an initial visual inspection, the CPD's are relatively similar in shape and range. The kinematics measures of velocity, acceleration, and jerk in all tool tip directions show the most similarity in shape. The force, distance from mean (\bar{d}), and the transverse and absolute tip directions measures show the most variability.

Segment 3 is the most different from the other two segments of the procedure, and this is seen in all tool tip directions. The axial forces are the largest in value, as this is a combination of the axial tip force and the grip forces not removed through the calibration process. This large axial force measure coincides with what was found for the expert surgeons (Kinnaird 2004).

The segments are then compared to a data lumping of the other two segments. These D-values and the corresponding confidence intervals signify the variability between segments (Figure 4.4). Each of the segments in an OR procedure represents a different portion of the dissection task.

The CPD reference (D_{RS-ref}) is created from resampling the reference CPD from itself. If the experimental D-values is close to the 95th percentile of the CPD (D_{RS-ref}), the more "similar" they are considered. The performance measure CPD's indicated the segments 1 and 2 are similar, and the D-value calculation verifies this.

As expected, segments 1 and 2 are more similar, and segment 3 is the more different from the other two segments. The D-values represent the variability between segments for this procedure. It gives us an idea of intraprocedural repeatability as each of the segments has a different goal in the OR, even though they are all considered part of the dissection task. In general, we can say that resident 1 is repeatable intraprocedurally.

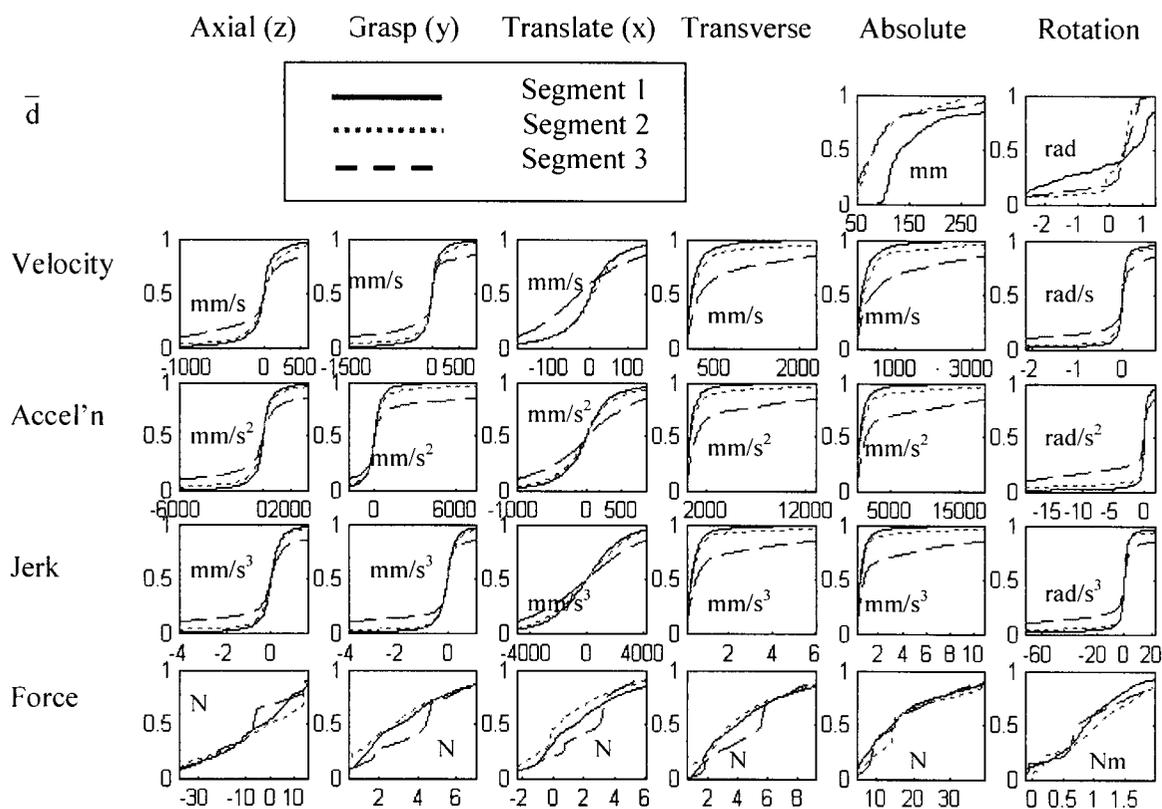


Figure 4.3: Resident 1 intraprocedure OR CPD. Segments 1, 2 & 3. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

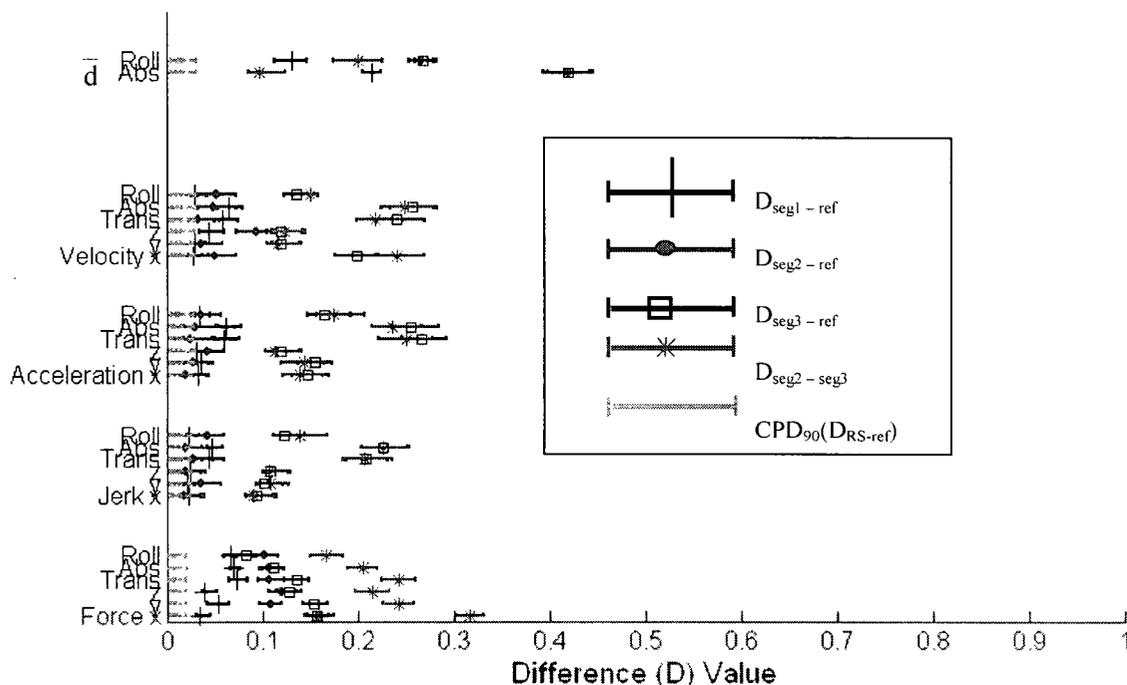


Figure 4.4: Resident 1 intraprocedure OR D-values. Segments 1, 2 & 3. The horizontal error bars represent the confidence interval on the D-value.

4.2.3.1.2 Resident 2: Intrasubject Intra-procedural OR

The CPD's of the calculated performance measures again seem to be quite similar in shape and range for all measures (Figure 4.5). We see some small differences in segment 3 in the transverse and absolute tool tip directions as was seen previously with Resident 1. Again, the force and \bar{d} show the most differences in CPD shape.

The D-values are calculated and lend support to what was seen in the CPD's of the performance measures (Figure 4.6). The kinematics performance measures have small differences between all segments with the majority of D-values below 0.3. We see here that segment 1 has many D-values that fall within the CPD ($D_{\text{RS-ref}}$) indicating the values are essentially the same. Also, for this subject, there are no D-values greater than 0.6. The kinematics performance measures all have D-values below 0.2 demonstrating very repeatable behaviour within this procedure.

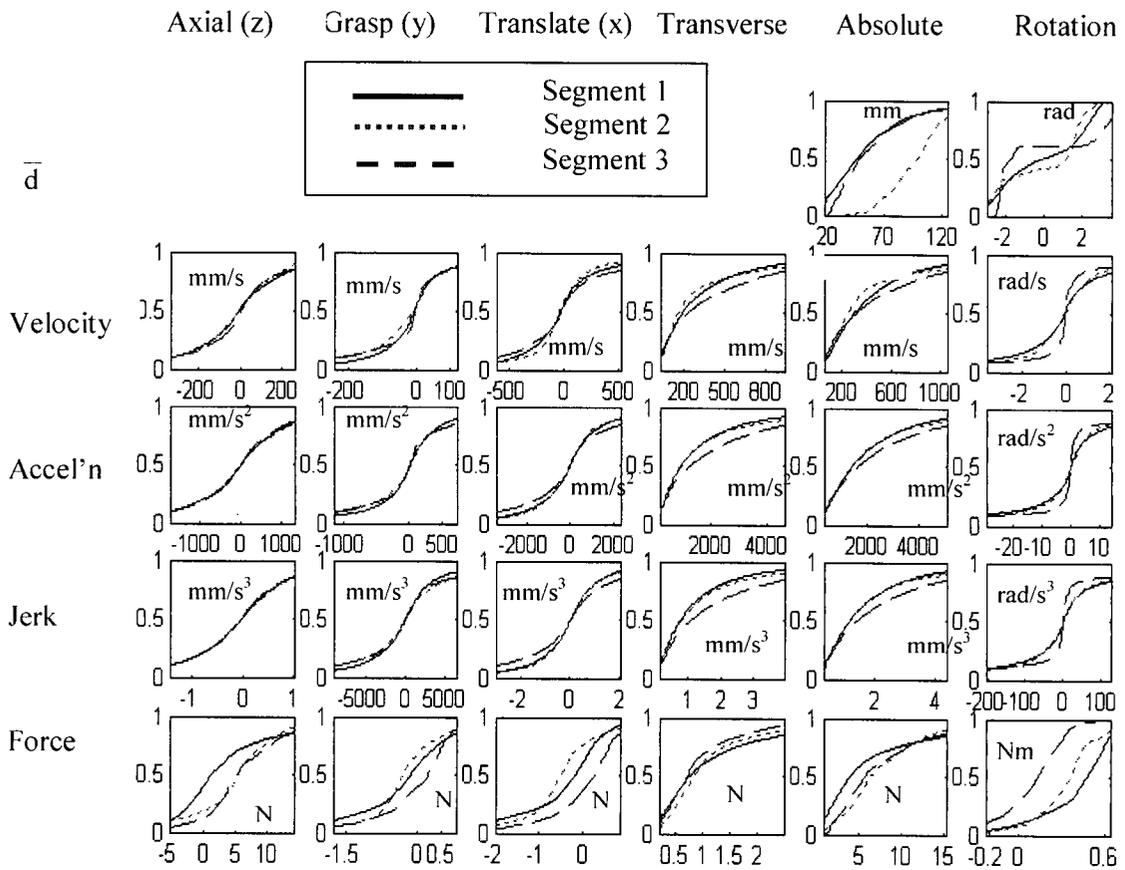


Figure 4.5: Resident 2 intraprocedure OR CPD. Ssegments 1, 2 & 3. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

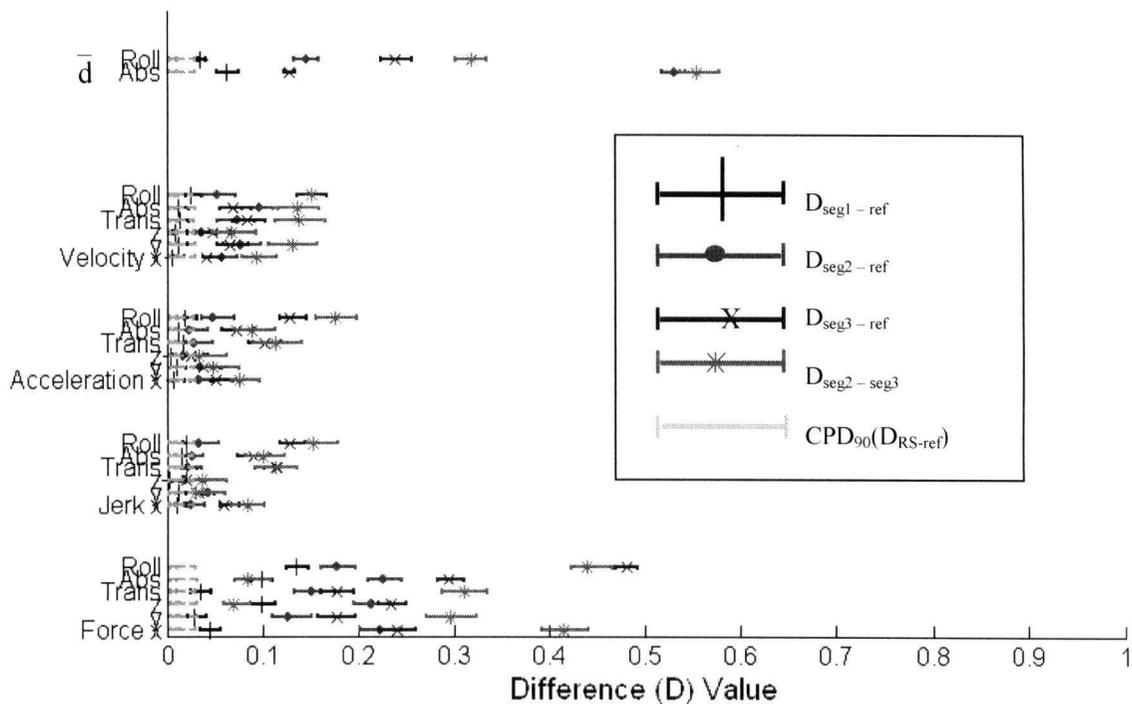


Figure 4.6: Resident 2 intraprocedure OR 2D-value. Segments 1, 2 & 3.

4.2.3.1.3 Resident 3: Intrasubject Intraoperative OR

For Resident 3, we see very similar results (Figure 4.7) as to what was seen with Resident 1 and Resident 2. The three segments show a lot of similarity when looking at the CPD performance measures. Segment 1 shows some difference in the jerk measure in the transverse and absolute tool tip directions. We again see the most difference in the \bar{d} and force measures.

The similarity between segments is confirmed by the D-values (Figure 4.8). Force and \bar{d} measures have again the largest differences. This OR trial shows the least amount of intersegment variability with all D-values below 0.3 except for the force in the translate (x) direction. Resident 3 demonstrated the most repeatable behaviour within a single OR procedure.

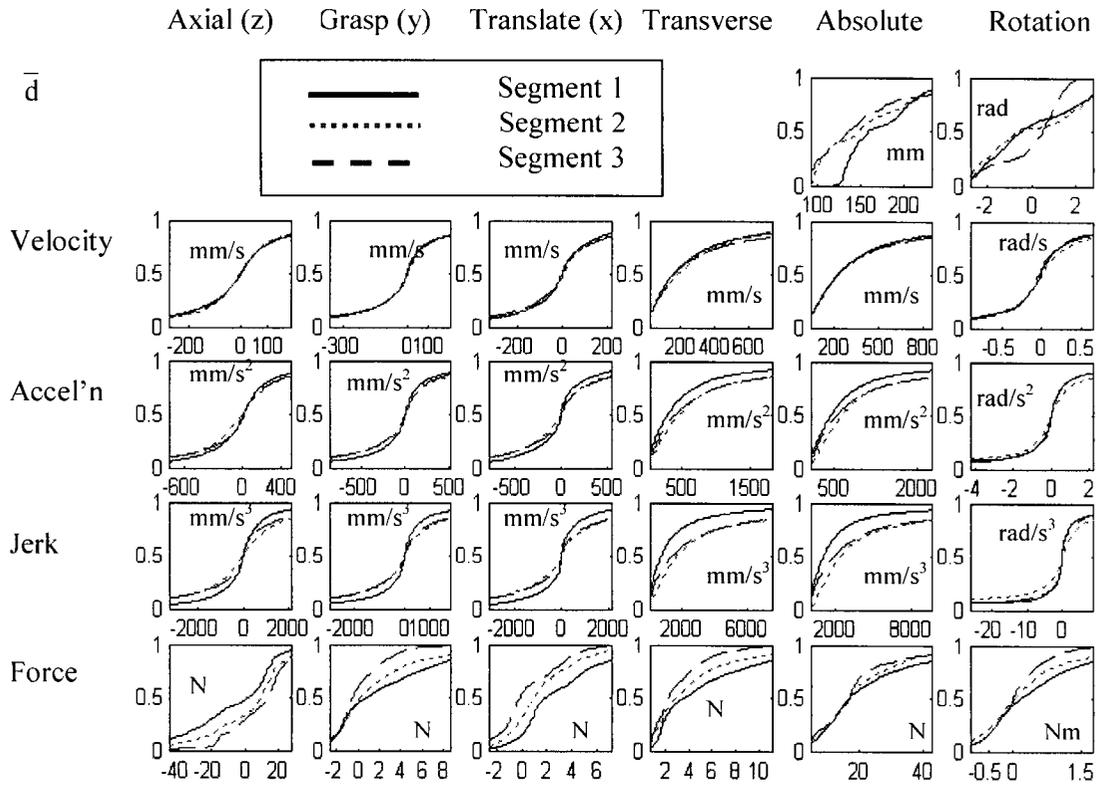


Figure 4.7: Resident 3 intraprocedure OR CPD. Segments 1, 2 & 3. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

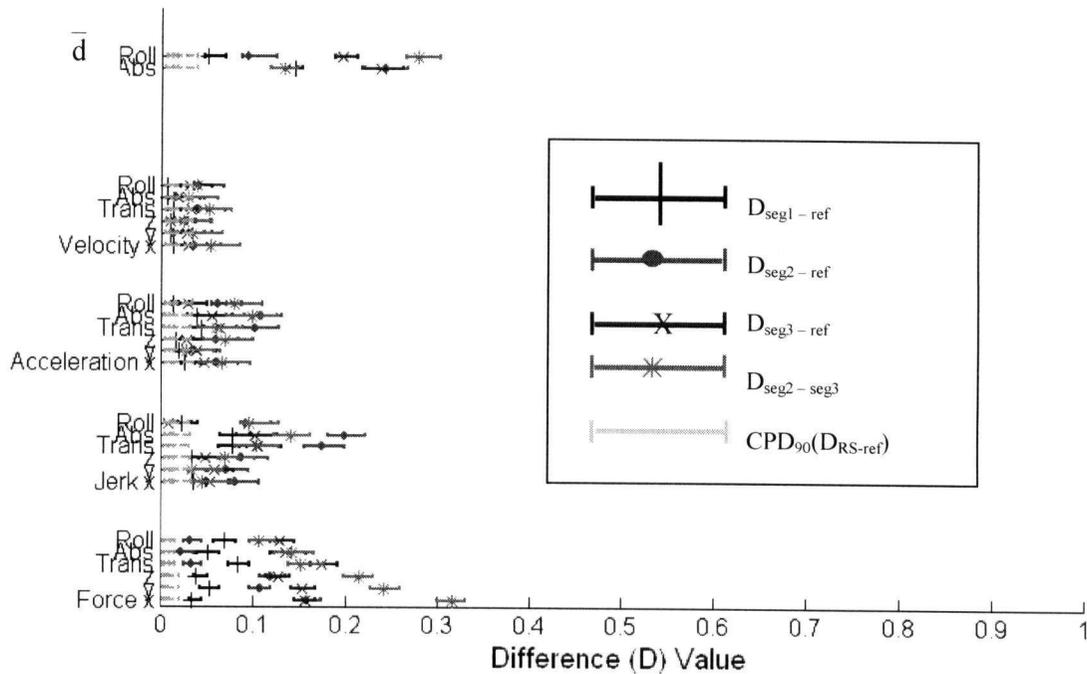


Figure 4.8: Resident 3 intraprocedure OR D-value. Segments 1, 2 & 3.

4.2.3.2 $\Delta 2$: Intrasubject Intertrial VR simulator

The three residents each performed the cystic duct dissection module on the VR simulator three times each. The performance measures extracted from the VR simulator are less comprehensive than from the OR or physical simulator; there are only 17 performance measures. There is not any roll direction, or component force values.

It should be noted that as of time of this manuscript, the VR force values are pending change. The manufacturer hardware calibration value was not quoted correctly, and therefore the VR force values will need to be multiplied by a factor still to be determined. We do know that this factor will be less than 2, and therefore will not significantly affect the comparison results.

Intrasubject intertrial variability was examined, and little variability was seen in either the range or shapes of the CPDs for all three residents (Figures 4.9, 4.11, 4.13). Also seen from the VR simulator data, are the low absolute force values and the small range of values. The residents also tended to spend about half of the time at very low forces.

The three trials D-values for each of the three residents were compared. These D-values (Figure 4.10, 4.12, 4.14) coincide with the visual observations seen on the CDF plots representing very little differences in the majority of measures. The largest differences are seen in the absolute force and distance from mean performance measures. The variability is so low in this contextual comparison that many D-values are below 0.1 between the three trials. Each of the three residents is very repeatable in three trials in the VR simulator.

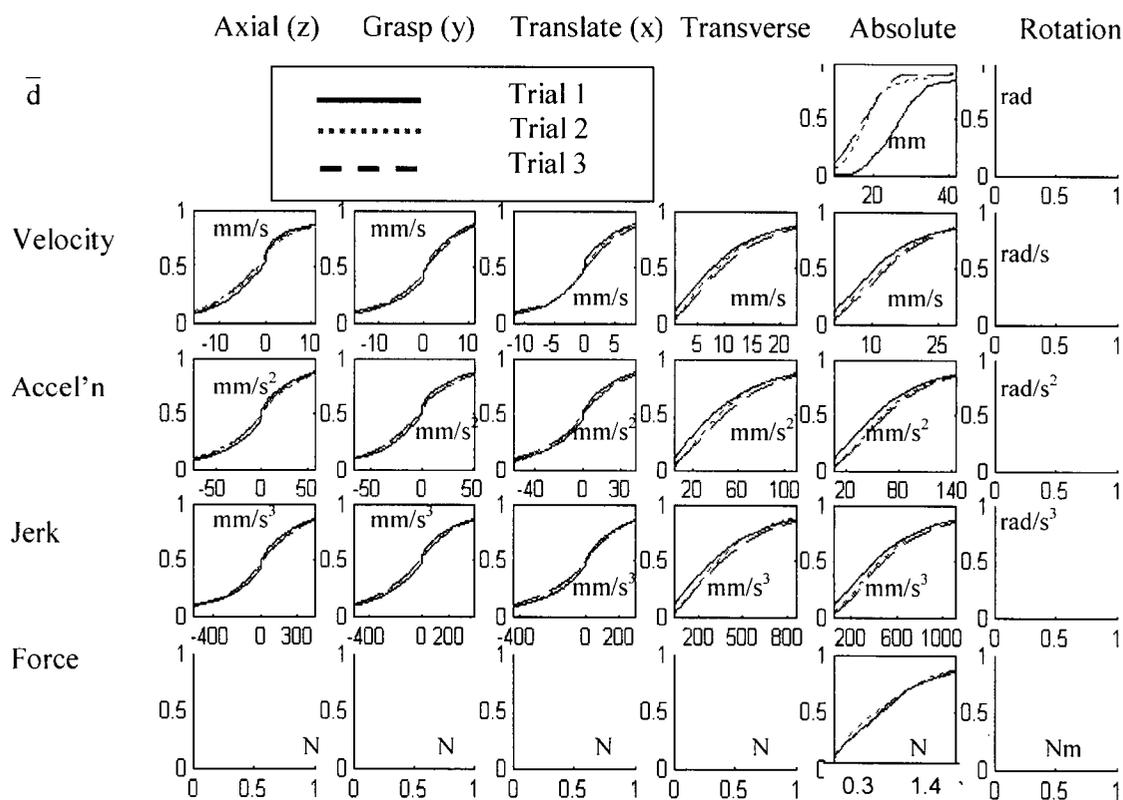


Figure 4.9: Resident 1 intertrial VR simulator CPD. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

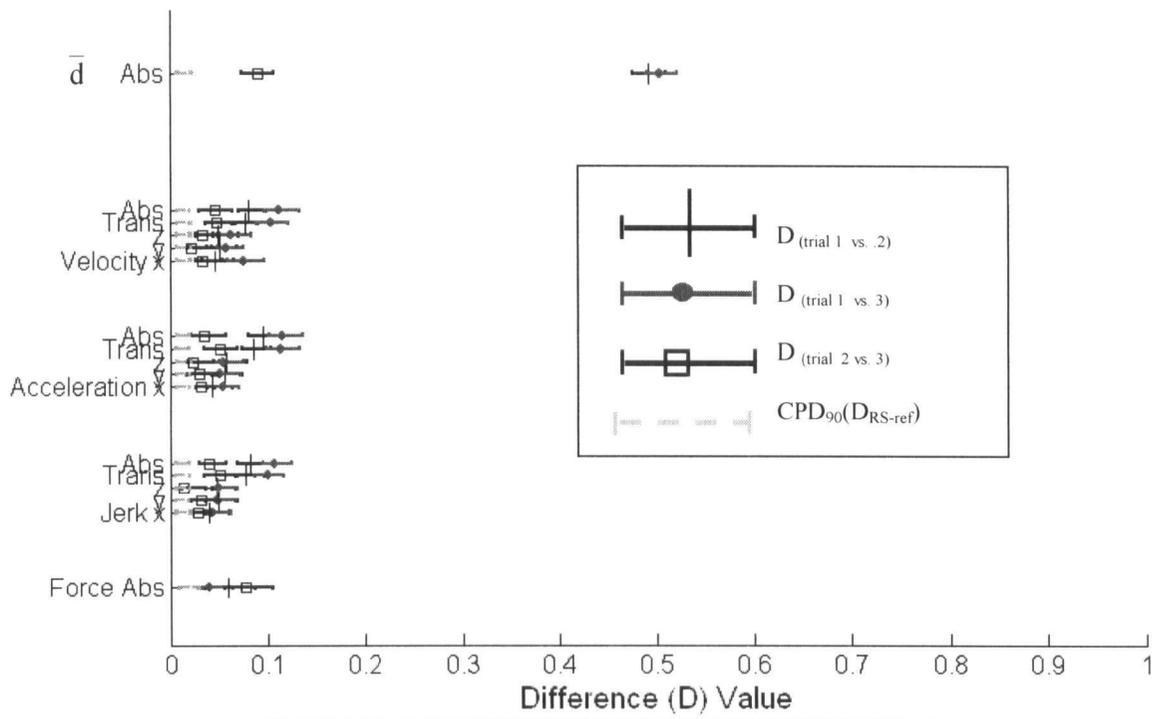


Figure 4.10: Resident 1 intertrial VR simulator D-value comparisons.

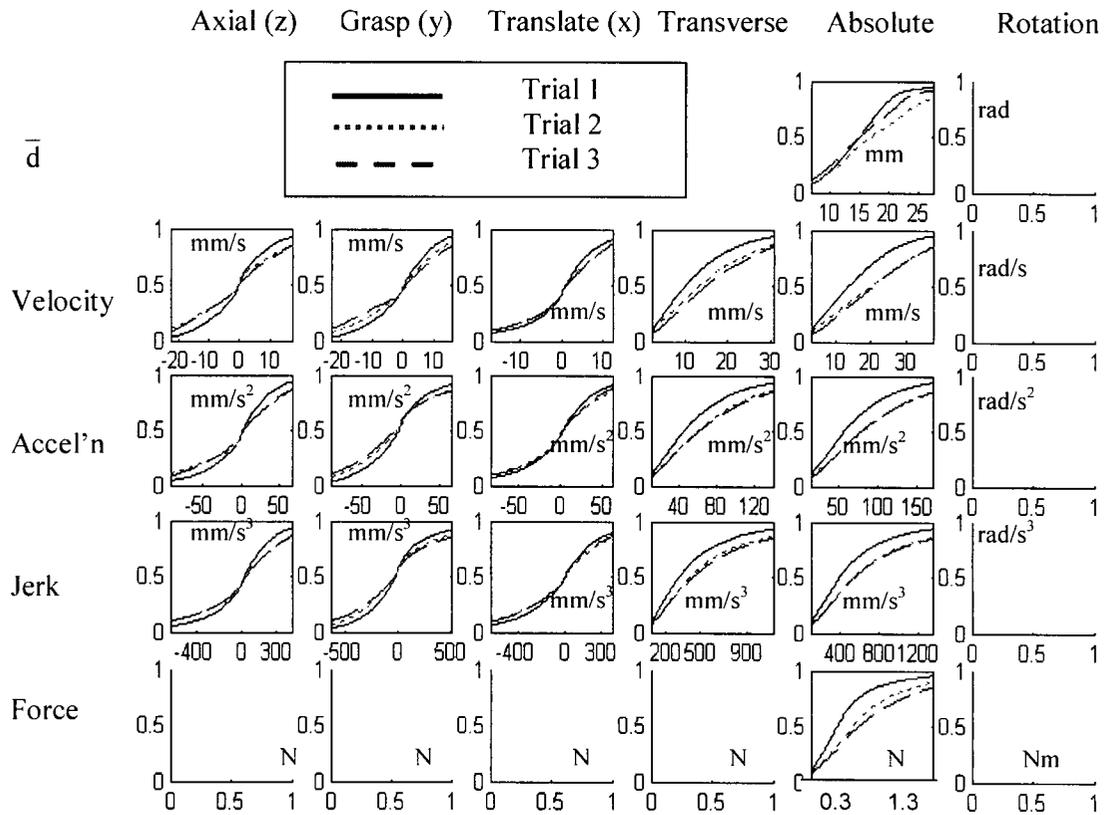


Figure 4.11: Resident 2 intertrial VR simulator CPD. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

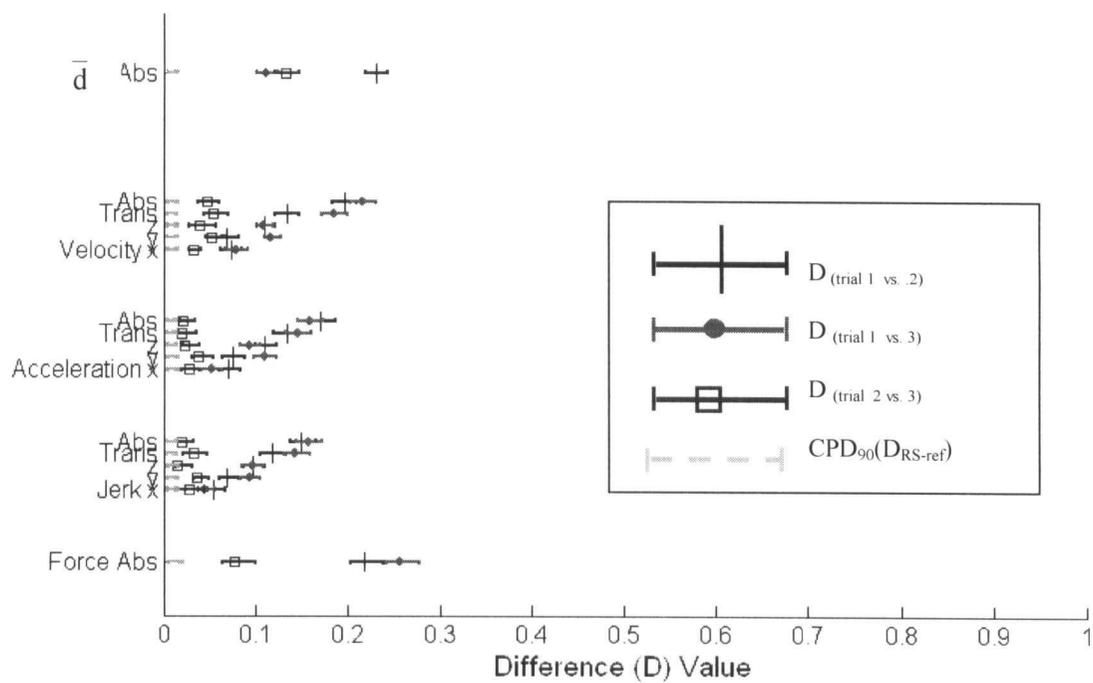


Figure 4.12: Resident 2 intertrial VR D-value comparisons.

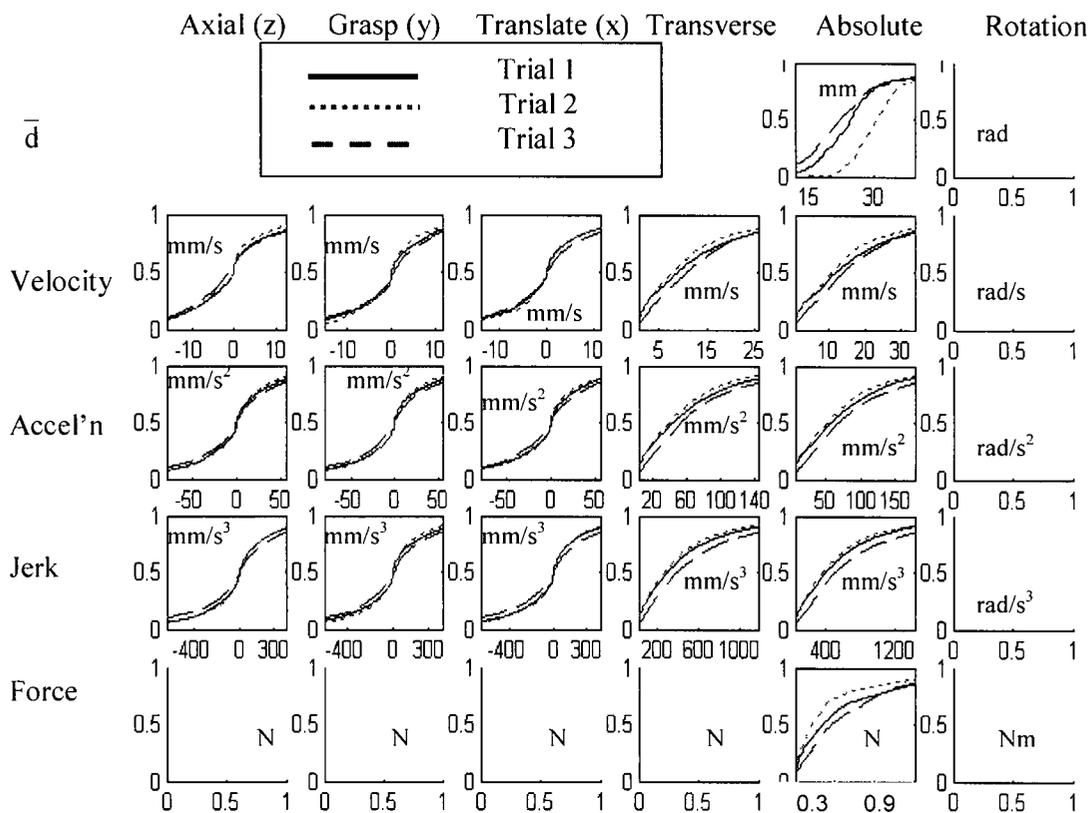


Figure 4.13: Resident 3 intertrial VR simulator CPD. Each of the individual graphs represents a performance measure in that particular direction at the tool tip.

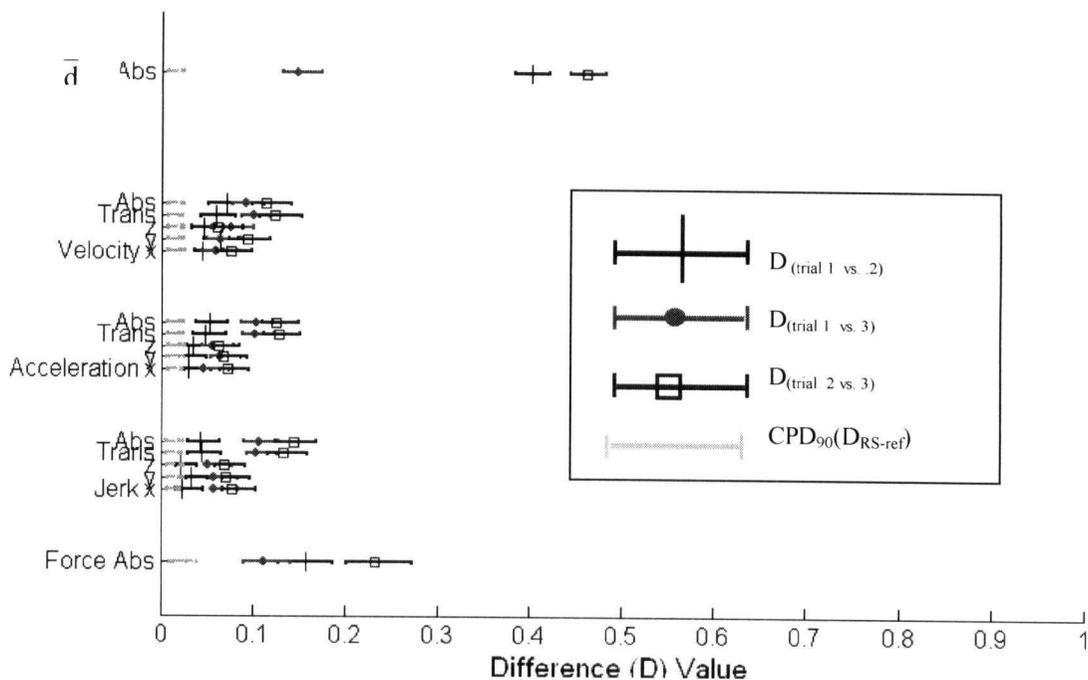


Figure 4.14: Resident 3 intertrial VR D-value comparisons.

4.2.3.3 $\Delta 3$, $\Delta 4$, and $\Delta 5$: Intersubject Intrasetting Comparisons

The three residents performance measure CPD's are compared in each of the three contexts of the operating room, virtual reality simulator, and physical simulator (Figures 4.15, 4.17, 4.19). We are able to examine consistency at the skill level by making these comparisons. At first glance, the CPD's for the three subjects in each context look rather similar. The operating room CPD's show the most differences, especially for Resident 2. The VR simulator CPD shows very similar shapes and ranges for all performance measures other than \bar{d} . The physical simulator CPD's shows again similar shapes and ranges in all measures other than forces and D mean.

The differences between the residents are analyzed (Figure 4.16, 4.18, 4.20). These D-values confirm the initial visual inspection of the CPD's. The data for the physical simulator shows that Resident 2 and Resident 3 have more similar patterns when compared to Resident 1 with most measure below $D = 0.3$. For the VR simulator, the three residents show much more similarity with the majority of the D-values below 0.15 demonstrating amazing consistency at their skill level. The OR difference comparisons indicate a larger range of D-values with a

spread throughout the range. A point to remember is that the VR simulator is a very repeatable and structured environment, while the OR context is much more inherently variable. And the physical simulator is somewhere in-between these two contexts in terms of variability and repeatable structure.

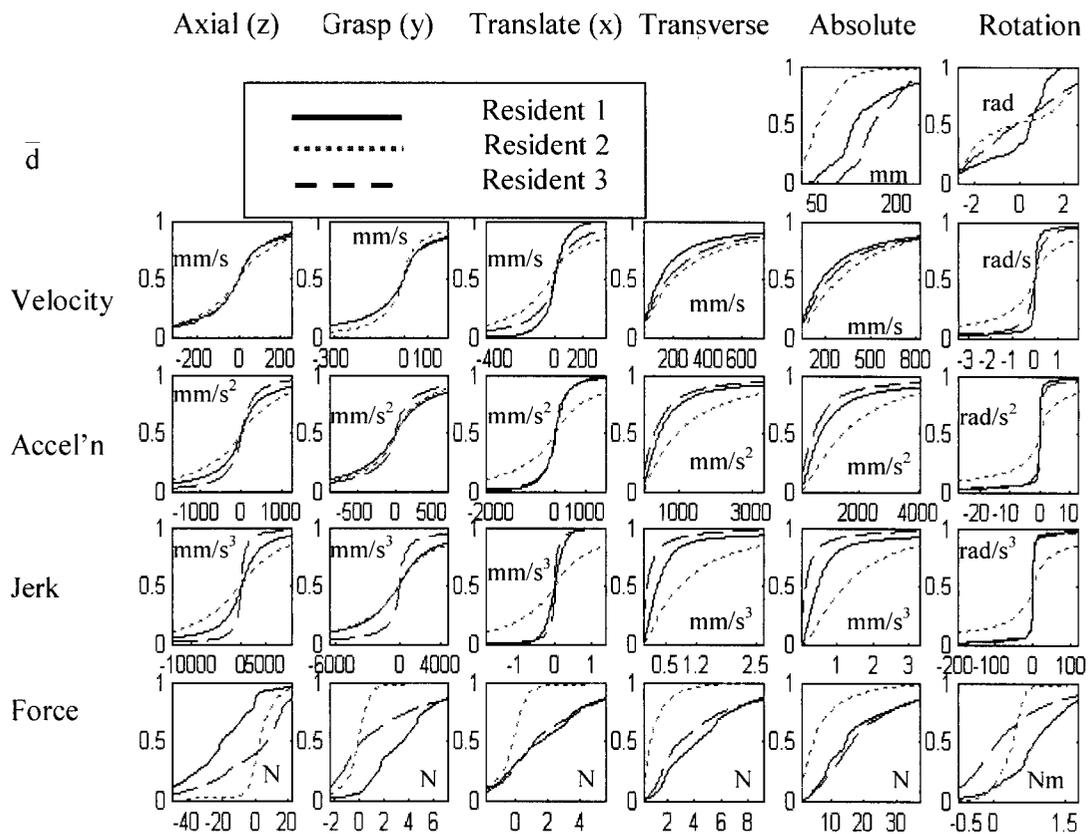


Figure 4.15: Intersubject intrasetting (OR) CPD.

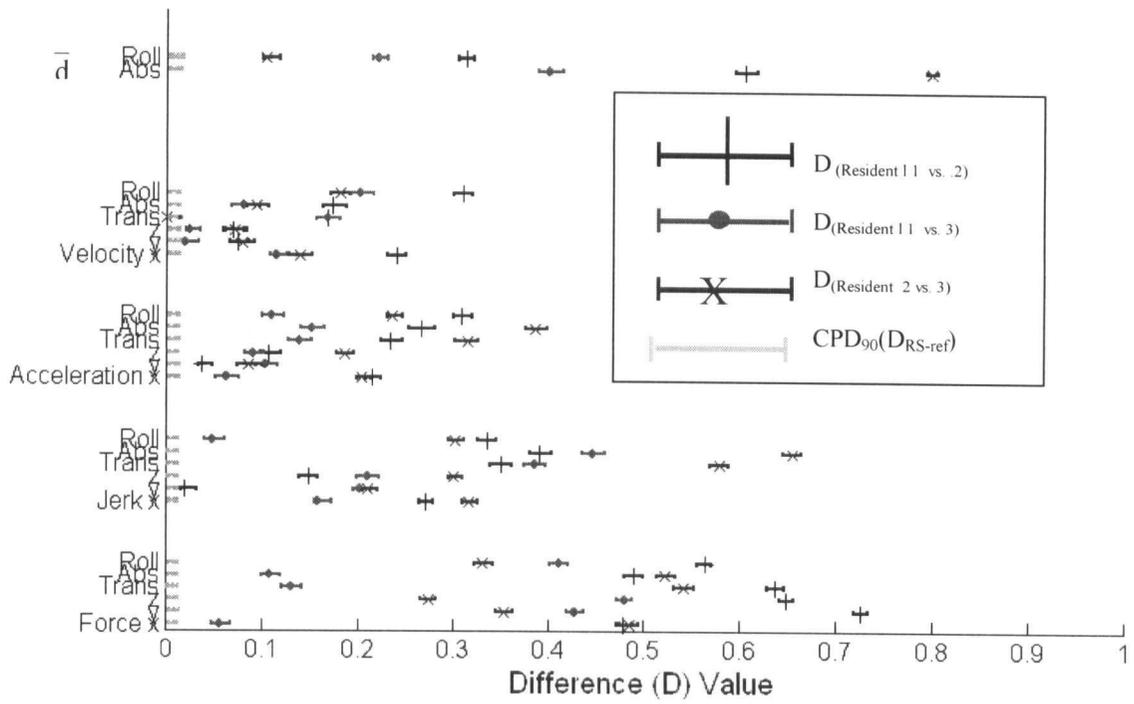


Figure 4.16: Intersubject intrasetting (OR) D-value comparisons.

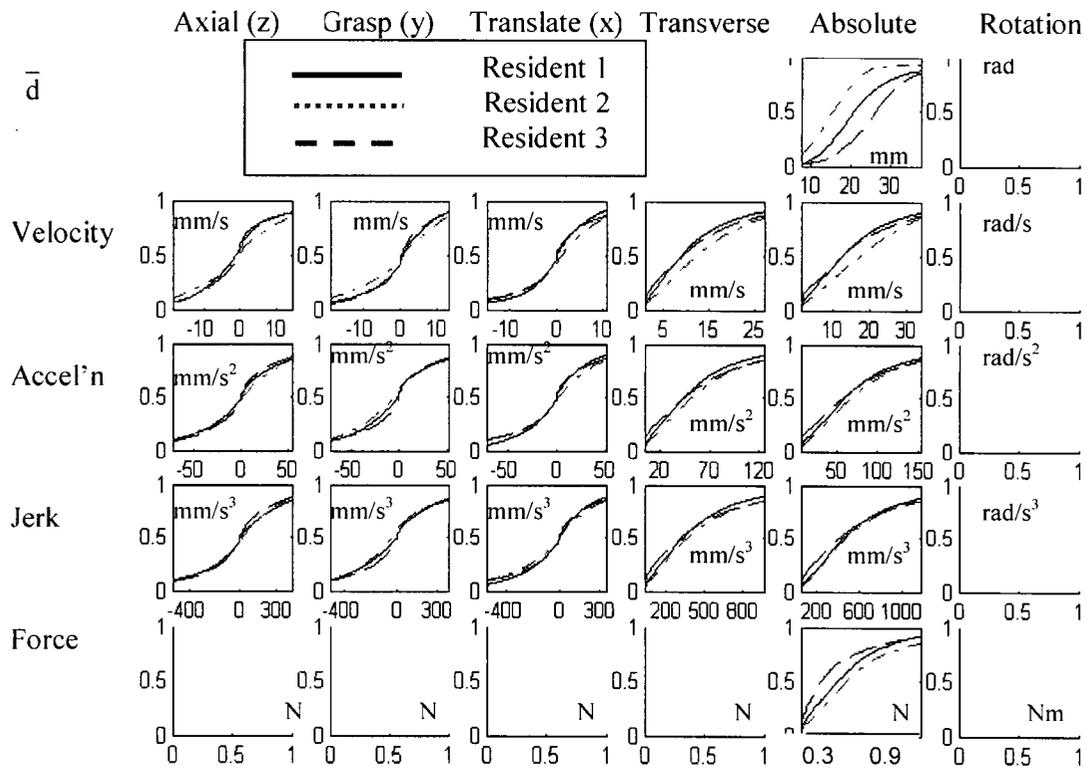


Figure 4.17: Intersubject intrasetting (VR simulator) CPD.

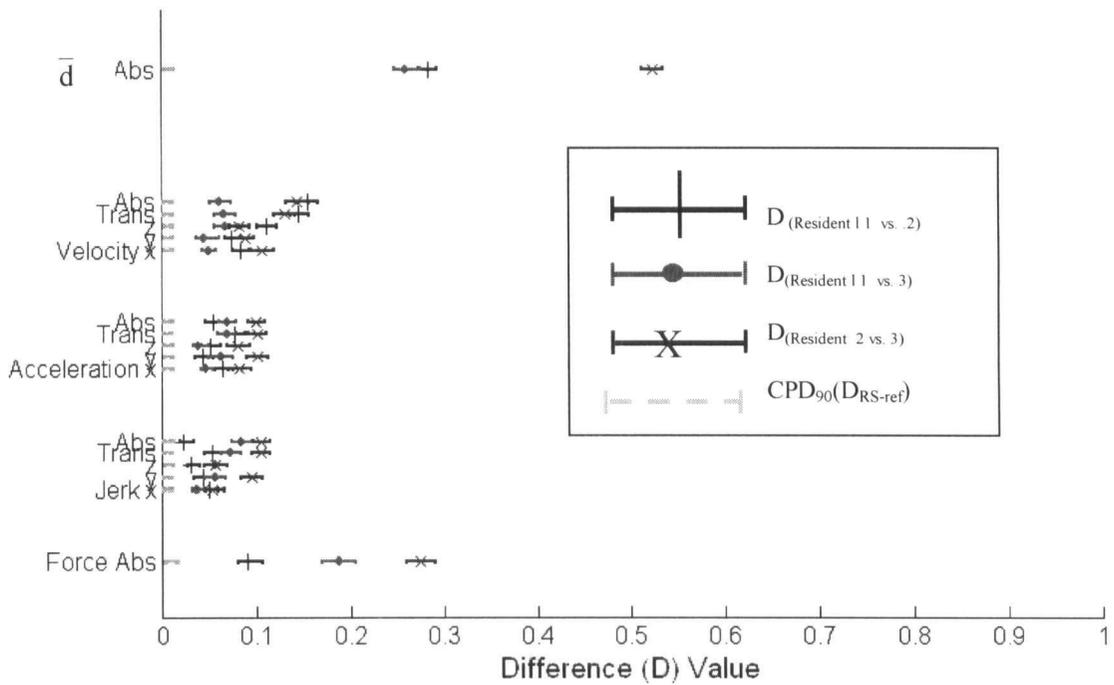


Figure 4.18: Intersubject intrasetting (VR simulator) D-value comparisons.

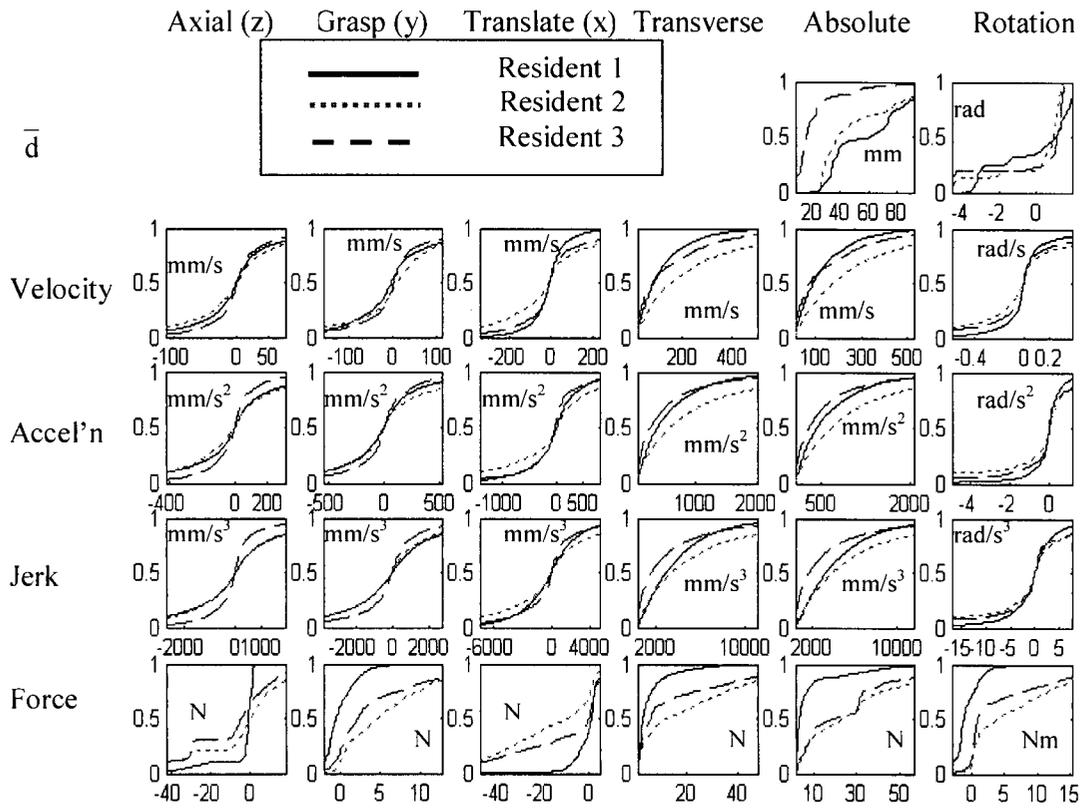


Figure 4.19: Intersubject intrasetting (physical simulator) CPD.

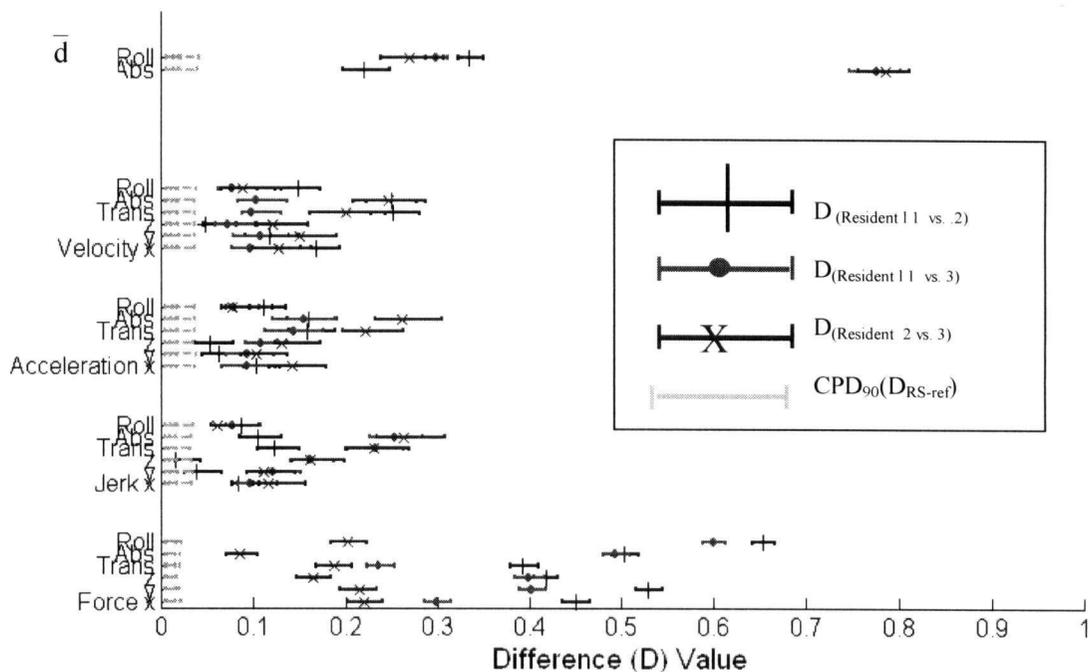


Figure 4.20: Intersubject intrasetting (physical simulator) *D*-value comparison.

4.2.3.4 $\Delta 6$: Intrasubject Intersetting

Each of the three residents had data collected in the three contexts: OR, VR simulator, and physical simulator. Comparisons were done for each subject in each of the settings (i.e. intrasubject intersetting) (Figures 4.21, 4.23, 4.25). These comparisons will help us in our investigation of the performance validity of the two surgical simulators. If the quantitative measures in the simulator are similar to that in the OR, then the simulator can be considered to show performance validity.

It can be seen from the CPD's that the kinematics measures in all tool tip directions for the OR, and the physical simulator are more similar when compared to the VR simulator. It would seem that the resident's move more slowly in the VR simulator relative to the physical simulator and in the OR. Another significant visual is the absolute force measure, which is very low in the VR simulator. It is so much lower than the physical simulator or OR settings that it is not easily seen on the plots.

The D-value analysis provides further evidence to the differences and similarities seen in the CPD's (Figure 4.22, 4.24, 4.26). The largest differences are seen between the force values, where the D-value is often 1.0, maximum absolute difference. Also in the \bar{d} performance measure, there are a few D-values that are also at 1.0. Another interesting note is that the comparison between the physical simulator and the OR, where many of the D-values are below 0.4. And conversely, when we compare the VR simulator to either the OR or physical simulator, the D-values are generally larger than 0.3. We consistently see that the OR vs. physical simulator comparisons shows lower D-values than the OR vs. VR simulator comparisons.

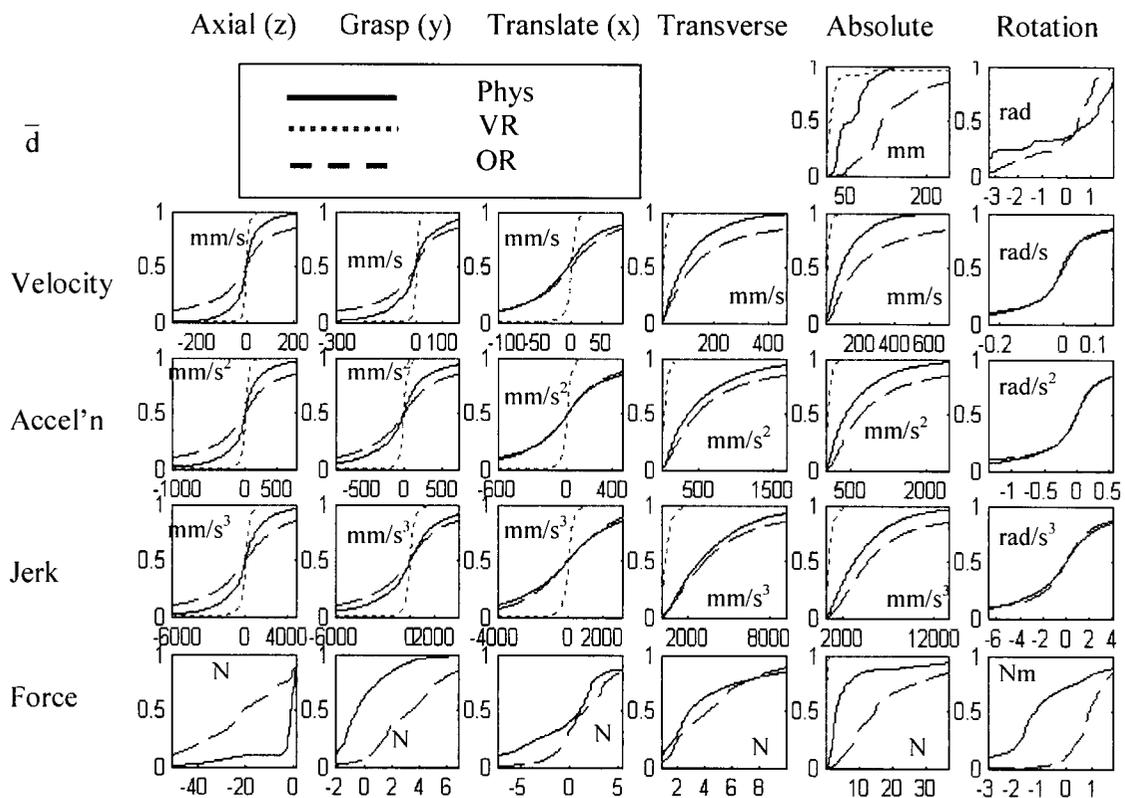


Figure 4.21: Resident 1 intersetting CPD. OR, VR simulator, physical simulator.

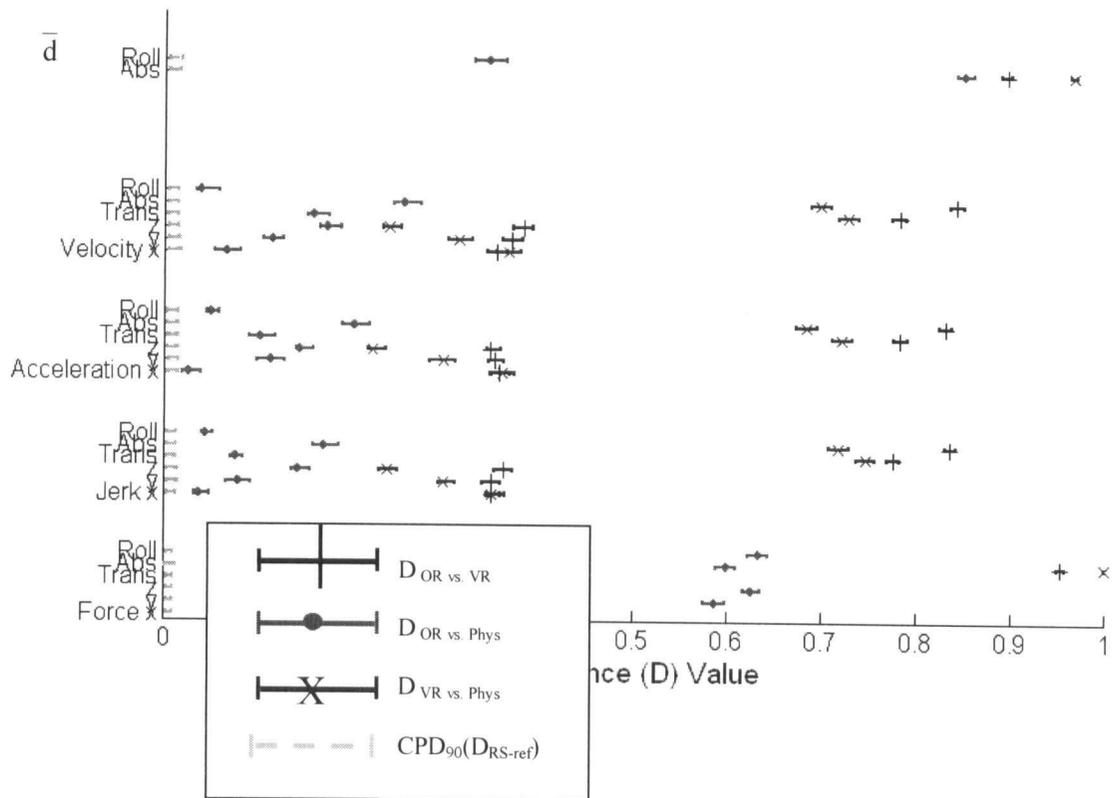


Figure 4.22: Resident 1 intersetting D-values. OR, VR simulator, physical simulator

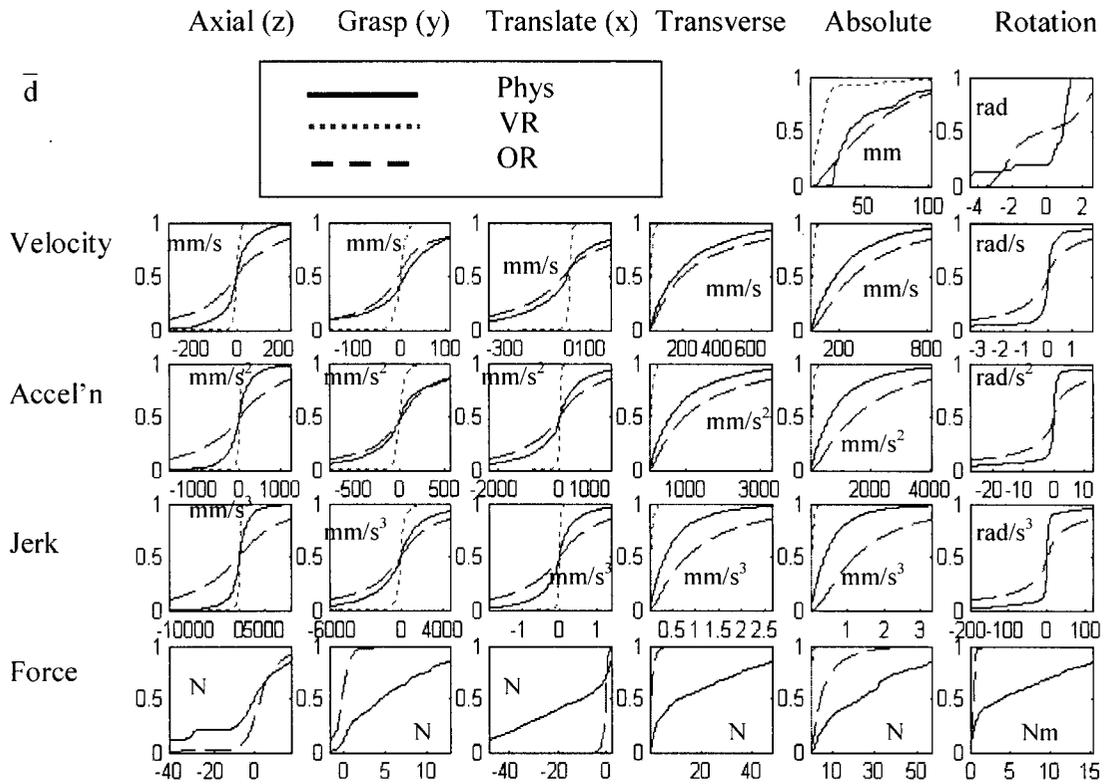


Figure 4.23: Resident 2 intersetting CPD. OR, VR simulator, physical simulator.

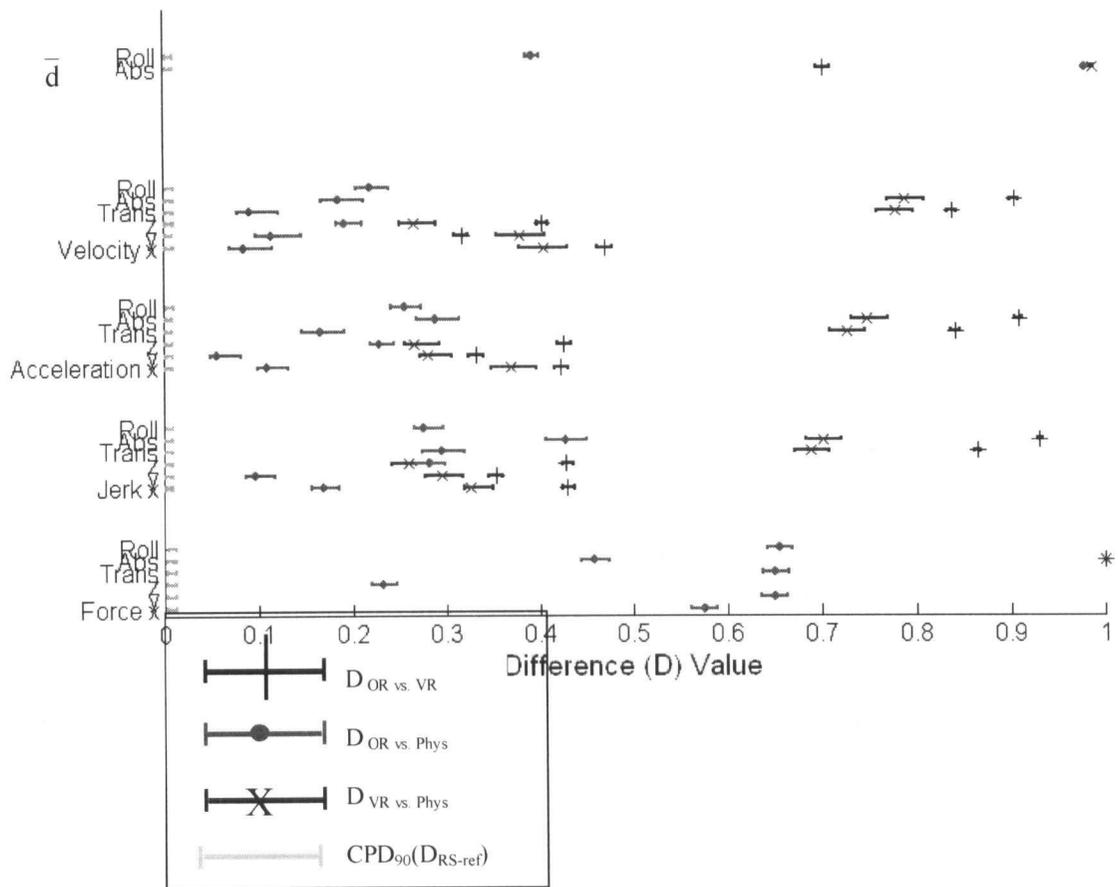


Figure 4.24: Resident 2 intersetting D-values. OR, VR simulator, physical simulator

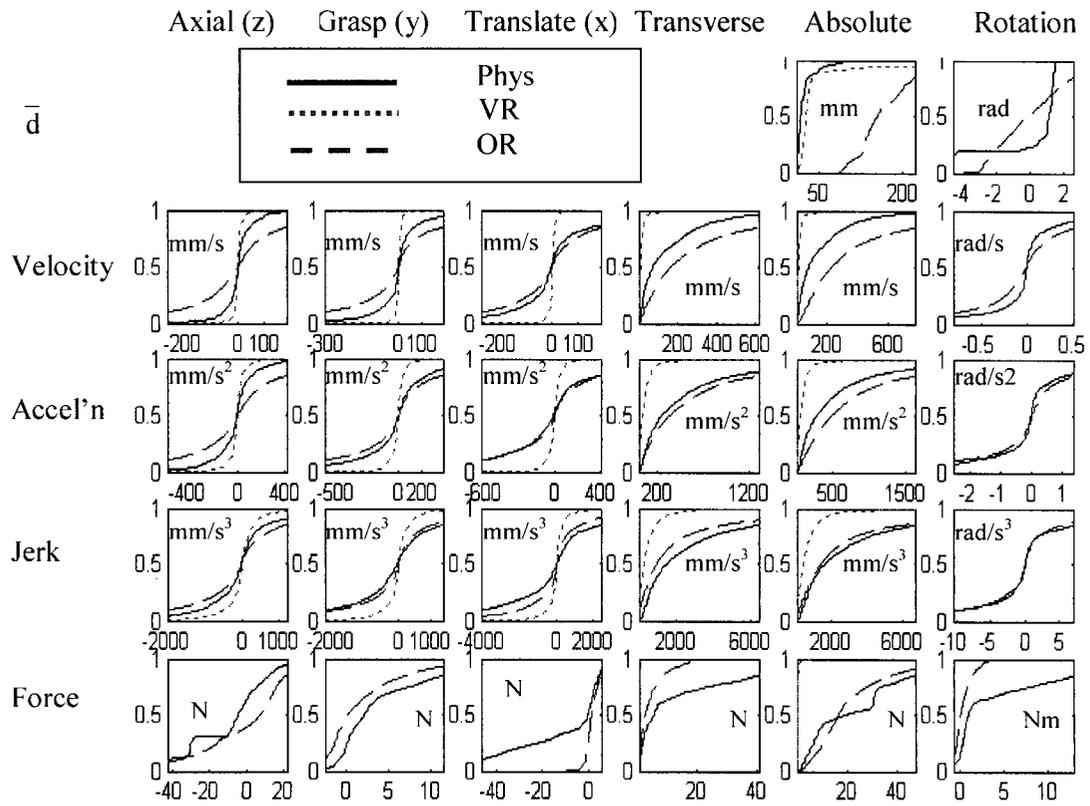


Figure 4.25: Resident 3 intersetting CPD. OR, VR simulator, physical simulator.

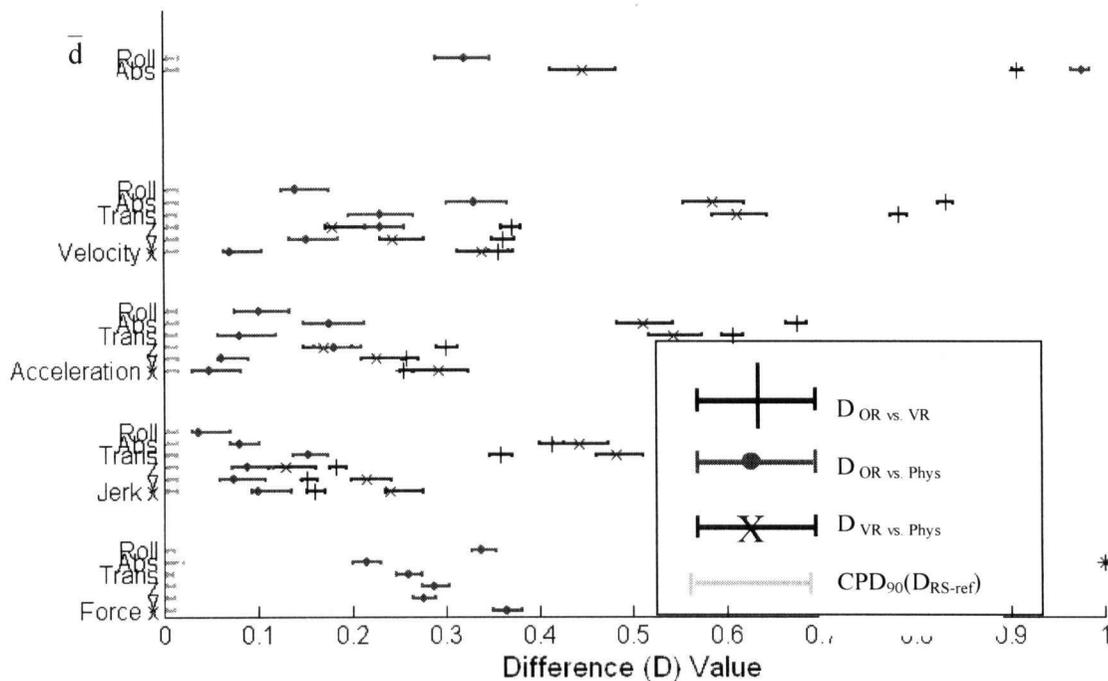


Figure 4.26: Resident 3 intersetting D-values. OR, VR simulator, physical simulator

4.2.3.5 Expert vs. Resident Comparisons

The data from the three residents was lumped together to create a large data set for resident surgeons in each of the three settings: OR, VR and physical simulators. The expert surgeon data (2 experts) collected and analyzed in a concurrent study by Catherine Kinnaird was also taken and lumped into a dataset to represent the expert surgeons (Kinnaird 2004). These two datasets in each setting, expert and residents respectively, could then be compared to each other to begin an investigation into the construct validity of the two simulators. If the simulator is able to detect skill level differences, it is said to show construct validity.

We are also able to demonstrate a new method for evaluating concurrent validity. This type of validity is usually assessed by a comparison to the “gold standard”, which is expert OR behaviour. This “gold standard” has been evaluated using checklists and rating scales in the OR. In our study, we are able to make the same assessments in all contexts, whether OR or simulators, or differing skill levels.

Due to intrasubject differences that cannot be clearly seen once the data has been lumped, and our small sample sizes for both experts and residents, we also investigated differences amongst the individuals. D-value comparisons are shown to analyze differences amongst the two experts and three residents. Each expert is compared to each resident individually for a more thorough construct validity investigation.

4.2.3.5.1 Interlevel Intrasetting OR

The performance measure CPD's for the lumped experts and residents were evaluated and plotted (Figure 4.27). The shapes of the kinematics measures of velocity, acceleration and jerk are somewhat similar, but the ranges do vary. We also see visual larger differences in the force and D mean CPD's.

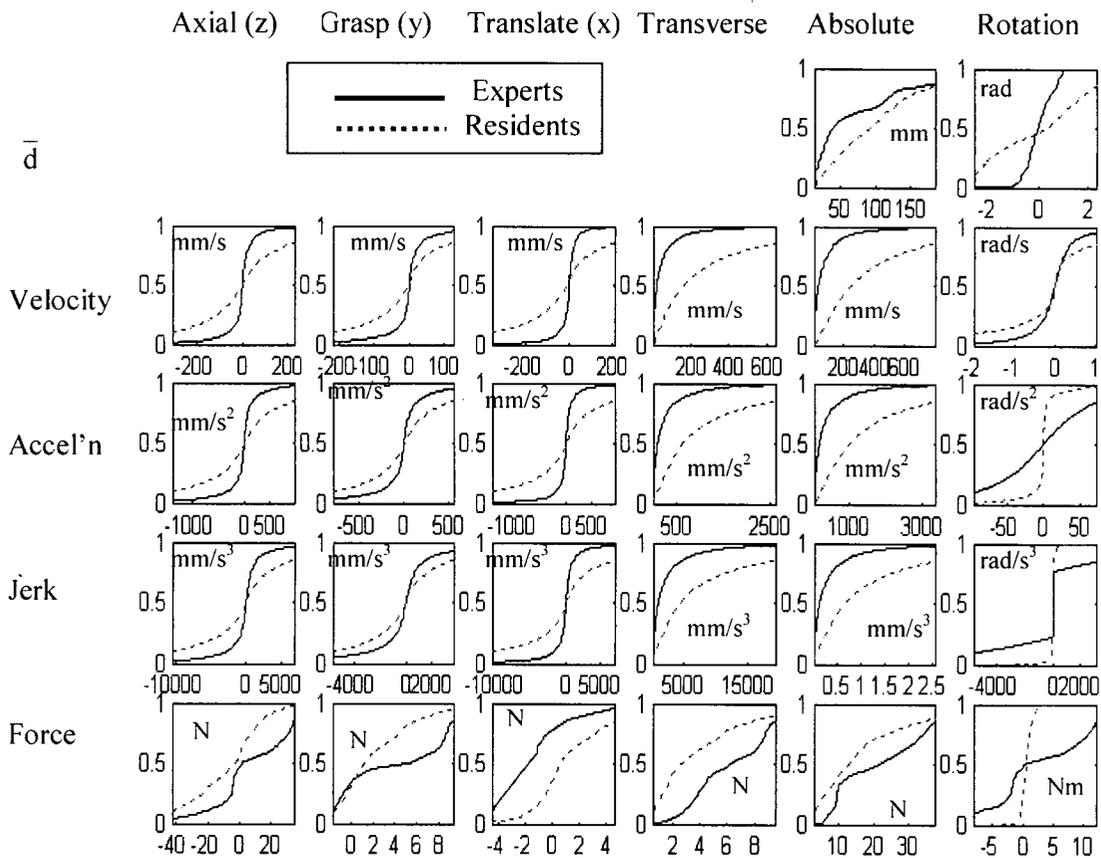


Figure 4.27: Lumped interlevel OR CPD.

We then investigate the individual differences for the two experts and three surgical residents (Figure 4.28). We see here the actual variation between all five subjects. We generally see similar shapes in the kinematics measures, while more variability in the \bar{d} and force measures.

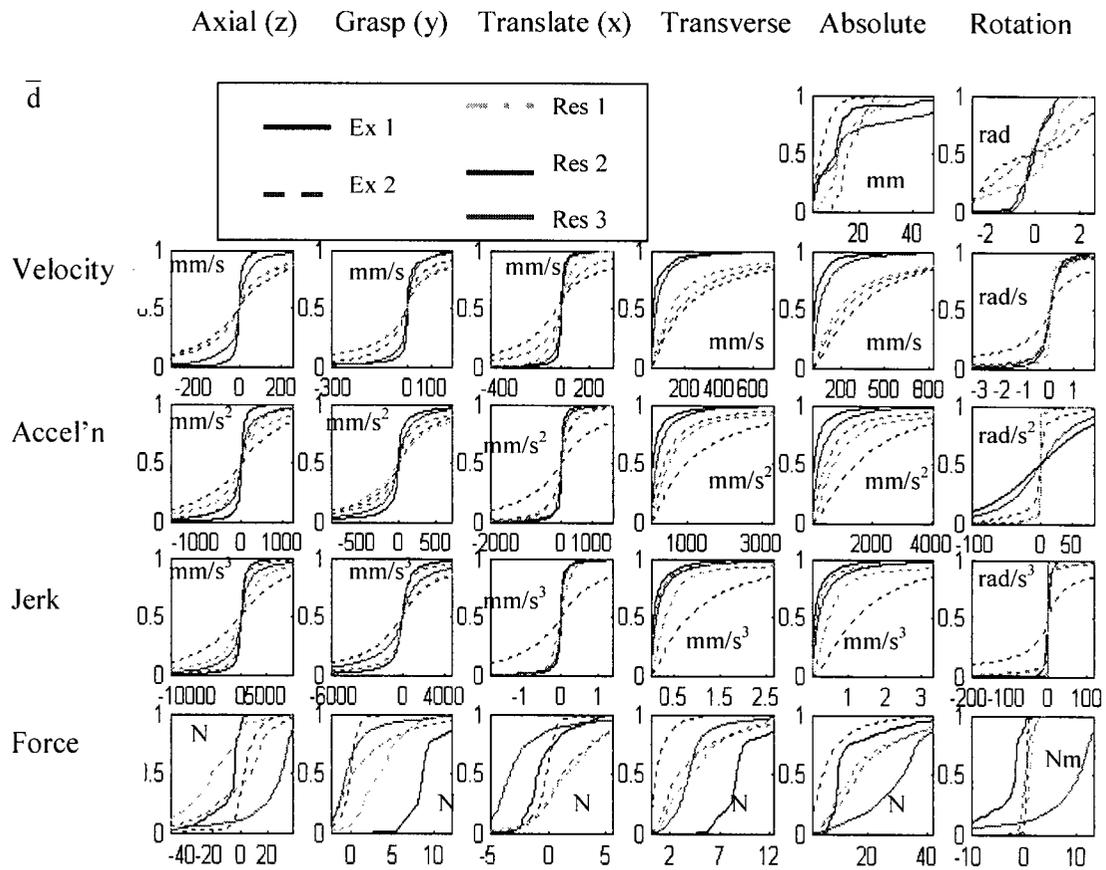


Figure 4.28: Interlevel OR individual CPD. Two experts and three residents.

An analysis of the D-values confirms what is seen in the CPD's (Figure 4.29). There is a wide range of D-values ranging from close to 0 to the maximum difference of 1. Again as was seen earlier in the resident comparisons, the force and D mean measures frequently have a D-value of 1. Here we see that expert 2 vs. resident 3 generally have D-values below 0.4, while expert 1 vs. resident 1 and resident 2 have all D-values greater than 0.2.

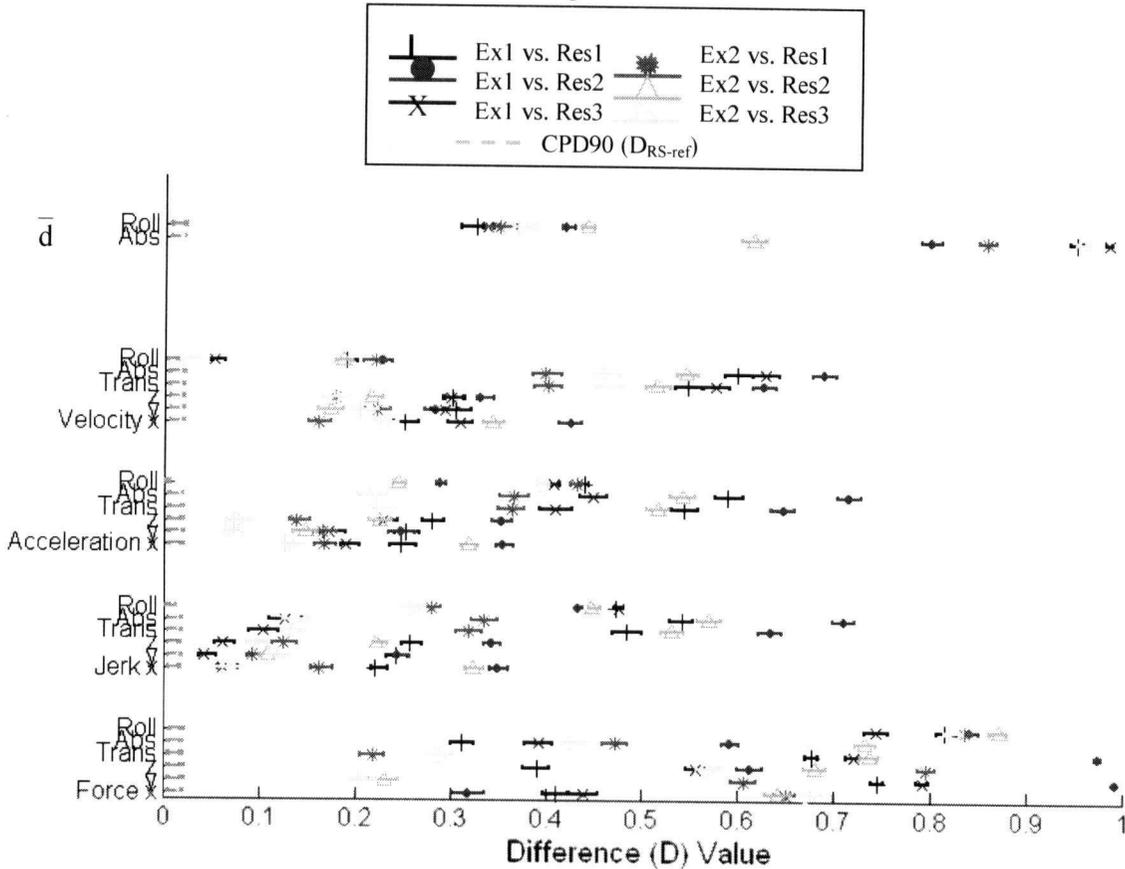


Figure 4.29: D-values for the two experts and three residents in the OR.

4.2.3.5.2 Interlevel Intrasetting Physical Simulator

Interlevel comparisons let us evaluate the construct validity of the physical simulator. We are looking for skill level differences. The CPD's of the interlevel physical simulator trials are shown in Figure 4.30. Here we see that the kinematics measures in all directions except roll seem to be relatively similar in shape and range. We again see the largest differences in the force and \bar{d} measures.

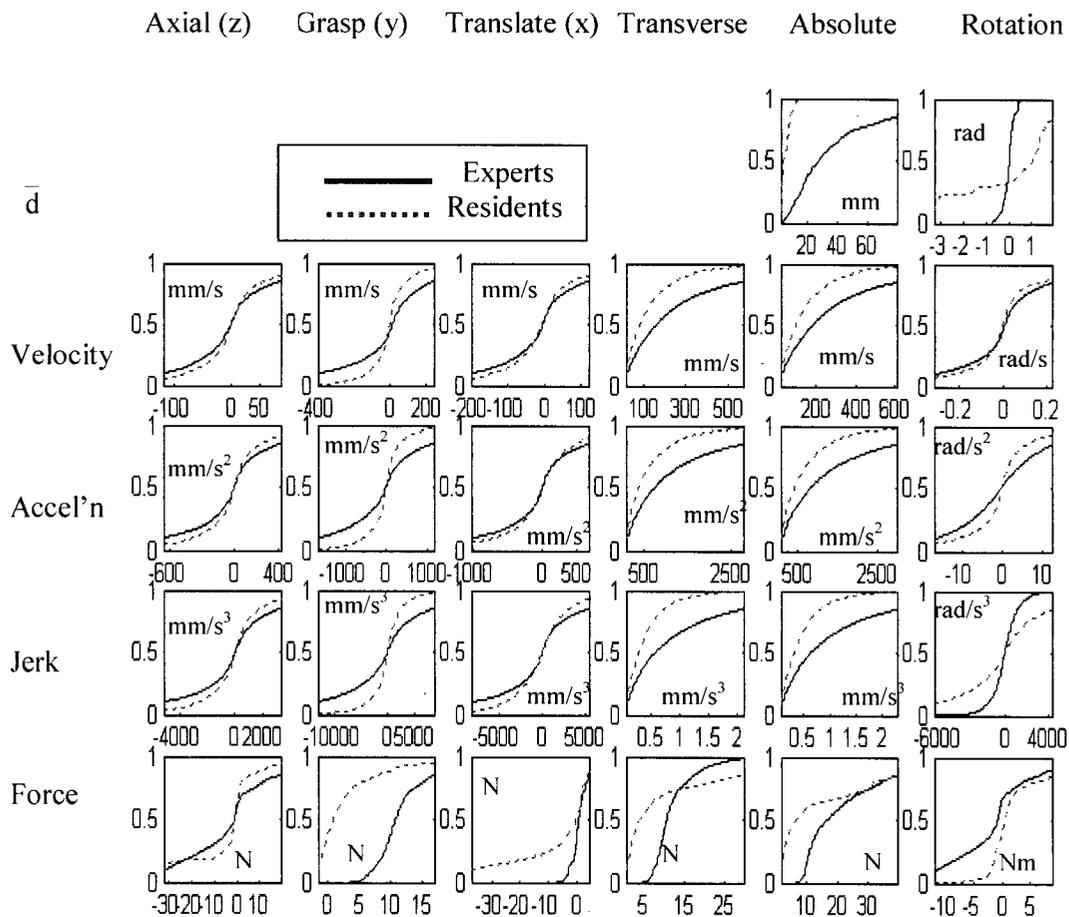


Figure 4.30: Lumped interlevel physical simulator CPD.

For the physical simulator, we again analyze the individual differences between each expert and resident (Figure 4.31). We do see general trends in the shape and range for the performance measures. We see slightly more similar CPD's than we saw in the OR comparisons.

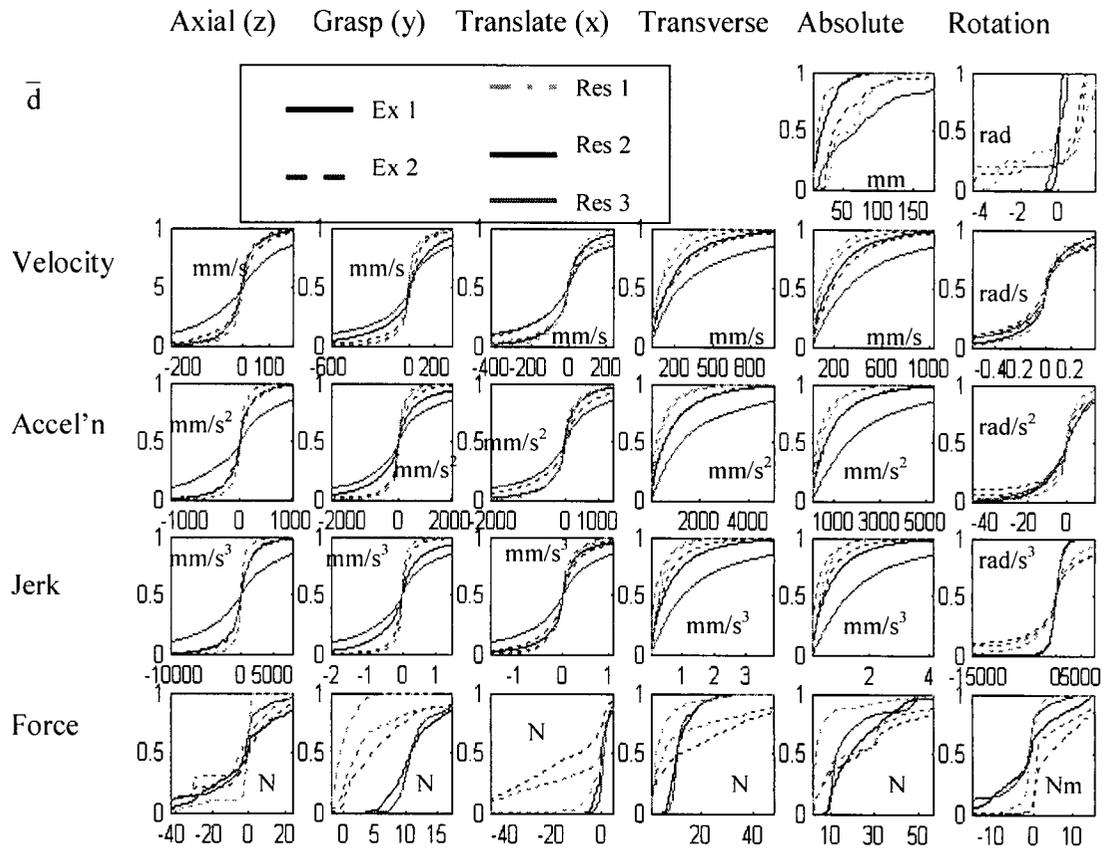


Figure 4.31: Interlevel physical simulator individual CPD.

By looking at the D-values for all the experts and residents, we can more clearly see the individual differences (Figure 4.32). There is a large spread of D-values throughout the range. And again, we see the largest differences with the D mean and force measures, with a few at the maximum difference of 1. It is also interesting to see that the comparison of expert 1 vs. resident 2, we get almost all the D-values below 0.2 indicating they are more similar in their behaviours.

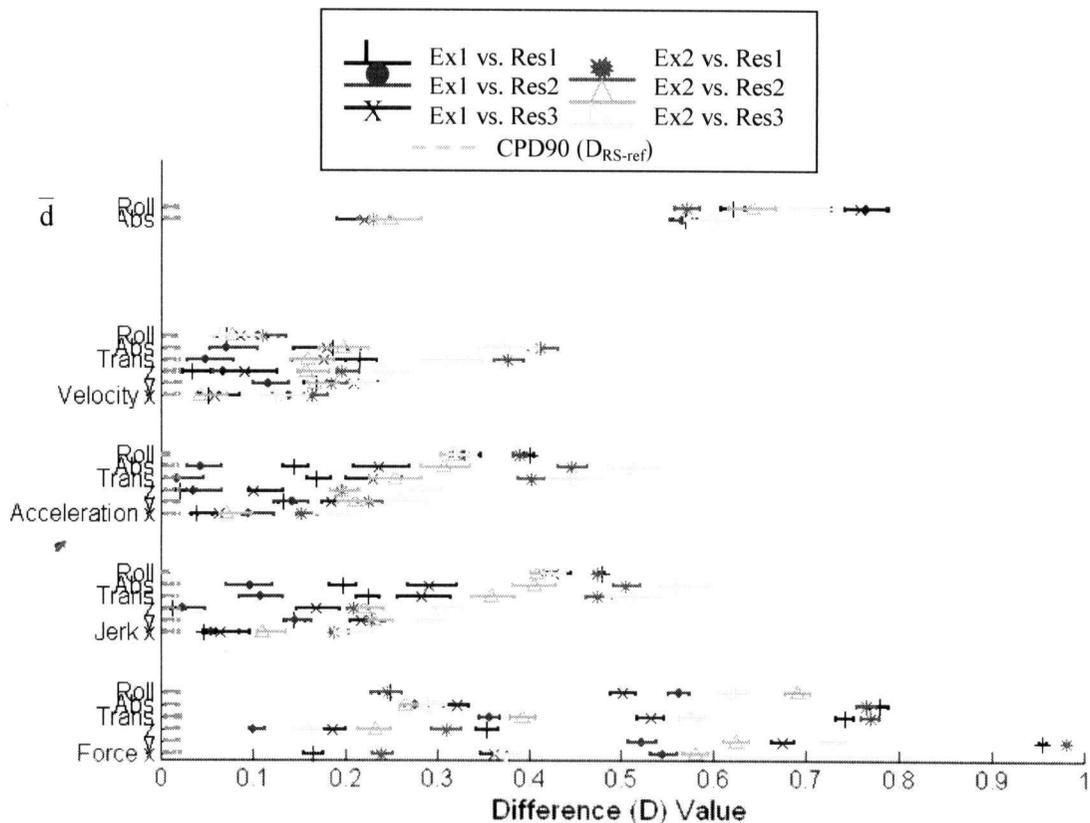


Figure 4.32: *D-values for the two experts and three residents in the physical simulator.*

4.2.3.5.3 Interlevel Intrasetting VR Simulator

Again, we are able to investigate construct validity of the VR simulator by looking for skill level differences. The CPD comparison between the experts and residents show the most similar profiles (Figure 4.33) when compared to the interlevel comparisons of the physical simulator and OR. The largest variations are seen in the \bar{d} and absolute force profiles. There are also differences in the transverse and absolute tool tip directions.

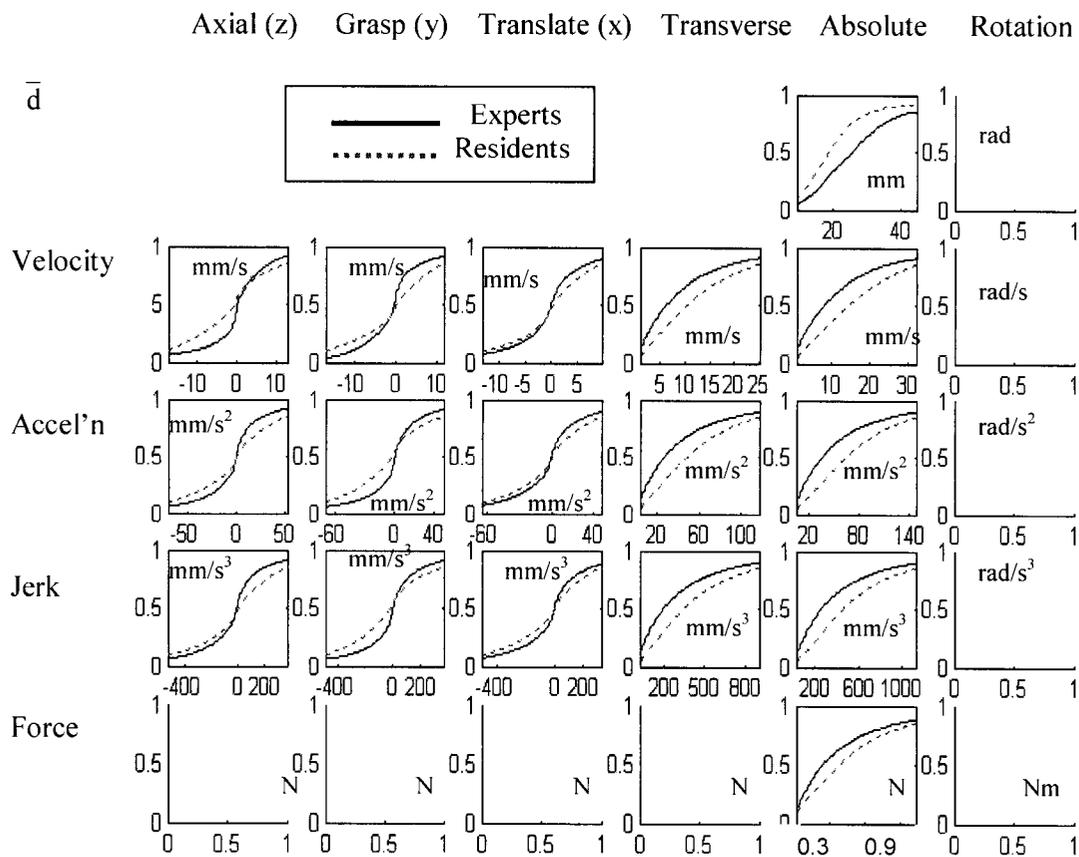


Figure 4.33: Lumped interlevel VR simulator CPD.

We then investigate the individual differences for the VR simulator for all residents and experts (Figure 4.34). Here we see a lot of similarity in all performance measures. The variability between experts and residents looks to be quite small according to the CPD.

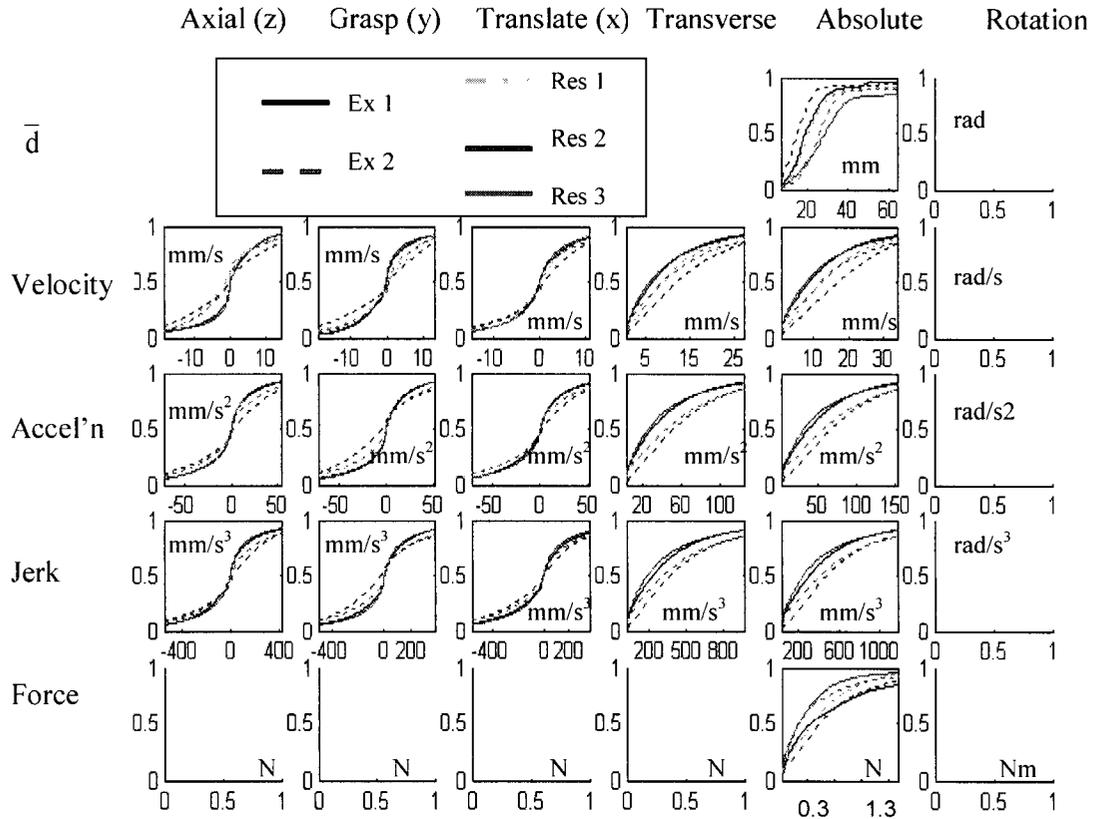


Figure 4.34: Interlevel VR simulator individual CPD.

The differences in the VR simulator are all much lower than what is seen in the physical simulator and in the OR (Figure 4.35). All D-values are below 0.4 except for the \bar{d} of expert 2 vs. resident 2. In this simulator, it would be difficult to distinguish between the experts and residents, as the differences are all small.

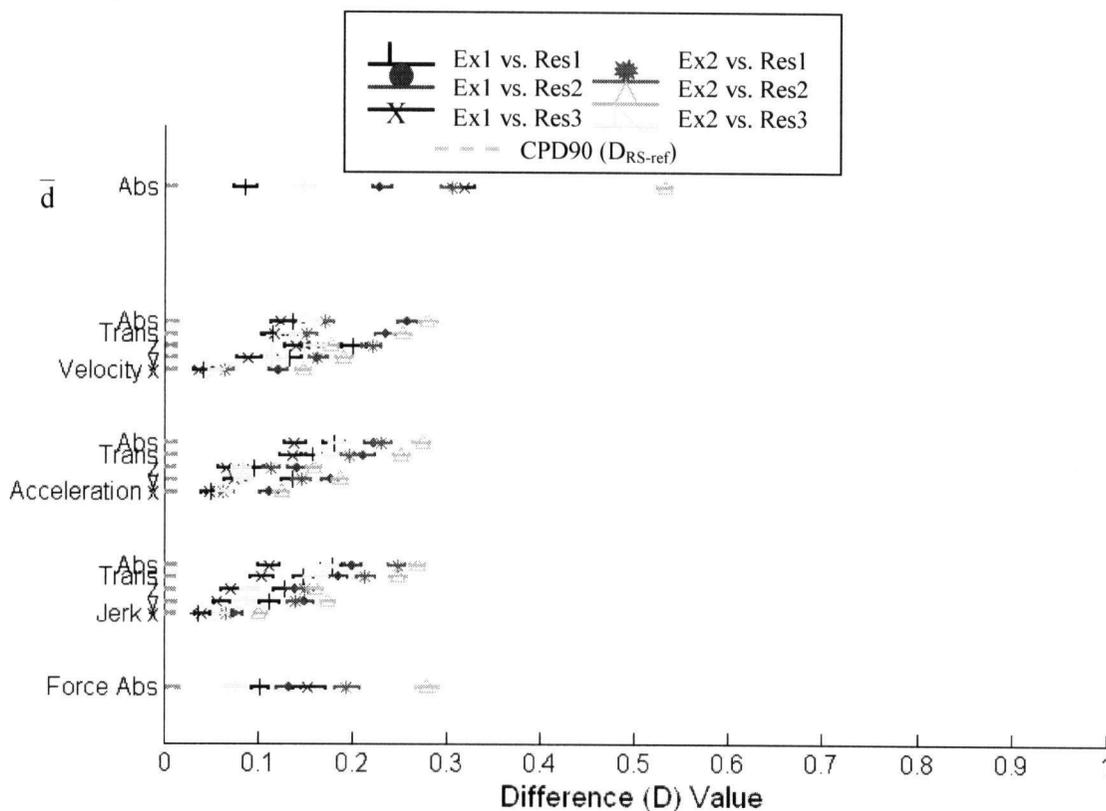


Figure 4.35: *D-values for the two experts and three residents in the VR simulator.*

4.3 Discussion

This project was chosen as a complementary and follow-up study to that of Kinnaird (2004). The results that have been obtained further support and answer the questions initially posed in Kinnaird's project. Further results in the realm of comparisons between expert and resident surgeons have also led to more questions and preliminary answers. This work begins the investigation of construct and performance validity of the physical and virtual reality (VR) simulators. We also have created a new method for evaluating concurrent validity by having the same performance measures in all contexts, with expert OR data as the "gold standard". The motor behaviour of the surgical tool tip was the model used to extract quantitative measures that allowed for comparisons. We will analyze and discuss the results to help us

understand the comparisons that have been made, and to further investigate the overall objectives of our validity studies.

4.3.1 Context Comparisons

4.3.1.1 Intraprocedural Operating Room Variability

The intraprocedural intrasubject operating room results show that there can be difference between the three dissection segments, but generally speaking, the three segments had a D-value of less than 0.3. This is an interesting result as each segment has a different objective (i.e. exploration, dissection, etc.). Each of these segments is a different section of a larger dissection task. The overall goal at the end of the dissection task is to have clipped and cut the cystic duct and artery to isolate the gallbladder, and this goal can be reached using a variety of kinematics and forces.

In two of the three trials, Segment 3 was found to have the largest differences from the other two segments. In the other trial, Segment 1 showed the most difference from segments 2 and 3. It is interesting to note here that in these three OR trials, the 3 segments chosen were not always of the same tasks as this was not possible.

Resident 1 spent the entire data collection period in the exploration and dissection phase, and we never were able to observe any clipping or cutting of the ducts/artery. Due to a chronically inflamed gallbladder, this exploration and dissection took more than 20minutes, when normally it would only take ~10 minutes. It is seen in the results that Segment 3 shows the most differences to the other two segments. It is possible that during this particular segment, some particularly difficult or different anatomy of the gallbladder was causing the resident to vary behaviours.

Resident 2 data showed the most difference in segment 3. This OR trial followed the “normal” segment protocol of exploration, setting and clipping the duct and artery, and a final gallbladder dissection segment. This protocol most closely follows that of Kinnaird (2004), but contrary to what she found, segment 3 showed the most difference as opposed to segment 1.

Resident 3 data followed the normal course of exploration, dissection, and the setting of two clips and cutting the cystic duct and artery. But segment 3 was the preparation of the cystic duct for cholangiogram examination (x-ray examination to look at gallbladder and ducts). The experimental surgical tool was used to insert the catheter for cystic duct exploration and verification. It is interesting to see that even though this is a completely different task than is done in any other OR trial, the performance measures do show some differences, but not as large as would be expected.

For the three OR trials, the levels of differences varied as expected. Resident 3 showed the least amount of intraprocedural variability with D-values below 0.3. While on the other hand, Resident 2 had the largest amount of variability with slightly more D-values over 0.3, while Resident 1 had D-values in-between these two. Generally, the three residents showed good repeatability in the OR intraprocedural comparisons.

4.3.1.2 Intrasubject Intertrial VR Variability

Intertrial variability in the VR simulator was found to be quite low with D-values in the ranges of close to 0 to 0.5, and most of the D-values were less than 0.2. This result was as predicted as the VR simulator is not an inherently variable situation. Also, all three VR trials were conducted consecutively on the same day. Each of the trials was the same scenario as the previous trials, and very predictable for the resident to know exactly what to expect. The results also coincide and verify the results by Kinnaird (2004) that a small number of trials are needed for each subject to study simulator performance. The largest intertrial differences were found for \bar{d} and the absolute force values, which was similar to what was found for the intraprocedural intrasubject OR trials. The three residents are very repeatable in their VR simulator performances.

4.3.1.3 Intersubject Intrasetting Comparisons

The intersubject intrasetting comparisons investigate consistency at the skill level within the context. We are specifically looking at PGY4 surgical residents.

Our results of intersubject intrasetting differences verify those found by Kinnaird (2004). The intersubject intrasetting differences decreased from OR to physical to VR. This result coincides

with the level of structure inherent in each context; least structured to most structured. The OR environment has many different variables that can lead to many differences, whereas the VR simulator environment does not have many variables, and is a predictable and repeatable environment.

4.3.1.3.1 Operating Room

The intersubject OR differences generally were in the area of 0.2-0.4 in all measures except for force where they were generally larger. This tells us that the residents will use relatively similar tool motor patterns to achieve the same end result. The force patterns used by the residents were more different, and again the same gall bladder removal procedure was completed successfully. The three residents show fair consistency in the OR context.

4.3.1.3.2 Virtual Reality Simulator

Intersubject VR simulator differences are lower than the OR trials. This is an expected result, as the intrasubject intertrial VR differences were very low also. The majority of D-values were below 0.1 showing incredible intersubject similarities. The three residents are very consistent to each other. It is an interesting result in that each of the residents received no training on the VR simulator, but would treat the simulator in a predictable and repeatable fashion to each other. The residents also commented that they thought the VR simulator was like a video game, and that certain tasks would be useful to train on a VR simulator. But this particular dissection task was not very realistic, and was not the same way they would behave in a real OR situation. These comments coincide with those of Kinnaird's experts' data comments on the face validity of the VR simulator (Kinnaird 2004). Neither residents nor experts felt that this VR dissection task was very good for training or evaluation of skills.

4.3.1.3.3 Physical Simulator

Intersubject physical simulator differences were also relatively low in all measures with most D-values below 0.3. These D-values fall in-between the OR and VR simulator differences, and this result is the same as found with Kinnaird's expert data (Kinnaird 2004). We see the largest differences in the force and \bar{d} difference values, and this is the same as was found for the OR trials' differences. It is also interesting to note that the comparison between resident 2 and

resident 3 resulted in all D-values below 0.3 except in \bar{d} . The three residents are fairly consistent in the physical simulator.

Again, a quick face validity study was conducted, the opinions varied amongst the residents'. Although none of them found the mandarin orange dissection incredibly realistic, their opinions did vary on how well they thought their motor patterns or force exertions were similar to in the OR. Another factor in the residents' opinion was the "juicy-ness" of the orange itself. Some mandarin oranges were quite juicy, and the skin did not peel off easily making for a more difficult dissection task. If the mandarin orange was generally "drier", the dissection task was easier, and the residents' were more easily able to complete the task.

4.3.1.4 Intrasubject Intersetting Comparison

The three residents were compared in the three environments of the OR, VR and physical simulators. This comparison gives us an indication of performance validity of the simulators, as we compare each to the OR environment. Specifically, these three contexts were compared to each other: OR to physical, OR to VR, VR to physical. These intersetting comparisons result in larger differences than the intrasetting comparisons. The D-values calculated run the entire range from close to 0 to 1 (similar to different). We see the largest differences between the VR simulator and both the OR and physical simulator settings. The most striking difference was between a few of the force measures of the VR and physical simulators with the three residents ($D=1$). All three residents show the similarity in differences in the kinematics measures where the VR simulator had slower velocities, accelerations and jerk measures when compared to the OR and physical simulator.

The most striking difference was between the absolute force measures of the VR simulator compared to the physical simulator and the OR, which is most visible when looking at the CPD. The residents did find and comment that the VR simulator to be a "low force" environment compared to a typical OR scenario.

4.3.1.5 Interlevel Intrasetting

By using the data collected and analysed in this project, and the data analyzed by Kinnaird (2004), we are able to begin an investigation into the construct validity of the VR and physical

simulators. A simulator showing construct validity will be able to detect differences between skill levels.

In this analysis, the data from the two experts was lumped together to create an “expert” group, and the three residents’ data was lumped together for a “resident” group. This is an efficient and easy method to detect immediate differences between the two skill levels.

We also looked at the differences between all 5 subjects, and can see how each of the three residents compared to the two expert surgeons. This is an interesting comparison as opposed to looking at the lumped data. We can see the more detailed differences between these groups.

4.3.1.5.1 Operating Room

Immediately on analysis, we can detect differences between the expert and resident data in the OR. Interestingly, the residents seem to be moving faster (velocity) than the experts. One would think that a surgical resident would be more tentative, and move slower, but as the data shows, this is not the case. We see large differences in the force data, where the expert surgeons use high forces more frequently than the residents. This could be a sign of the tentativeness of the residents. They may not feel comfortable in the OR to “pull and tug” with a lot of force.

When we look at the three residents and two experts individually, we see the differences cover the entire range from close to 0 to 1 (similar to different). We do see that the force measures are $0.2 < D < 10$. This tells us that the experts compared to the residents use different force patterns when in the OR. In the end, the same end result is reached, but the method to reach that point does vary significantly. The kinematics measures do tend to stay below 0.6, which indicates some more similarities in these motor behaviour patterns.

4.3.1.5.2 Virtual Reality Simulator

The interlevel intrasetting differences in the VR simulator are the smallest of the three contexts. In this context, it would be more difficult to make a conclusion on construct validity of the VR simulator, as the D values are small (< 0.3).

When we compare the subjects individually, we see very little difference in all measures. Both expert 1 and expert 2 show similar kinematics and force patterns to all three residents. This is a significant result as we are trying to detect differences between experts and residents. And according to this, we do not see significant differences between the skill levels. Therefore, the VR simulator does not pass the construct validity test.

4.3.1.5.3 Physical Simulator

Here in the physical simulator, we have interlevel difference levels in-between what was seen in the OR and VR simulator contexts. The physical simulator is more able to detect the differences between the two skill levels. We do see more differences in the force and \bar{d} measures, as was a common theme in all our context comparisons. In the individual comparisons, our physical simulator is the “middle of the road” setting, where differences are between that of the OR and VR simulator. The physical simulator can detect the skill level differences in a fair manner.

4.3.1.5.4 Experts vs. Residents

Now that we have collected and analysed data from both surgical experts and residents, and made some comparisons, can we conclude that if a resident behaves like an expert that they must be an expert? Being able to perform the same tool motor behaviours as an expert does not necessarily make you an expert. Our expert surgeons have been practicing surgery for many years, while the surgical residents are just at the beginning of their careers. So there must be other factors that determine whether a resident is of an expert’s calibre. Possibilities that could be studied include linking behaviour and outcome. Some of these outcome measures could include: surgical complications, mortality, loss of function, recovery time, and post-operative pain. Another study could be surgical errors, where surgeons could be doing the same behaviours, but one has more errors, and therefore increasing the risk. Our study has given insight into the motor behaviour patterns of the experts and residents, but we have not investigated the outcomes. These types of outcome studies could provide further evidence on what determines an expert.

4.3.2 Performance Measure Reliability

The results found in this study further the reliability of our chosen performance measures. Our intrasetting kinematics and force measures, especially in the VR simulator, were very consistent. We also see similar, although not to the same degree, consistencies in the OR and physical simulator settings.

As was first noted by Kinnaird in the study of expert surgeons (Kinnaird 2004), the force performance measure showed the most variability in intersubject and intrasetting comparisons. The results presented here agree with this, and further support the fact that the force measure is sensitive. The distance from mean measures also showed larger variability than the other measures.

4.4 Conclusions

Using the hybrid experimental tool and data collection system, we were able to successfully collect data from the human OR, and the VR and physical simulators for three surgical residents. The KS statistic (D-value) was used to make comparisons between settings and subjects to quantitatively assess motor behaviour, and simulator validity. The reliability of our performance measures was shown by low variability in the intraprocedural intrasubject comparisons. We also saw low variability in the VR intertrial comparisons for all three residents. The VR simulation is a very repeatable environment, and our performance measures also agree with this repeatability. Our intrasubject intersetting (OR, VR & physical simulators) showed much larger differences suggesting poor performance validity of the VR simulator, as the residents do not treat this context similarly. The physical simulator suggested an indication of fair performance validity as it was treated more similarly to the OR by all three residents. We also investigated interlevel differences to study the construct validity of the simulators. Some differences were noted between the skill levels (expert and resident), so it can be suggested that the physical simulator showed fair construct validity. The VR simulator differences were very small, so it would be difficult to conclude that it also shows construct validity.

With our limited sample sizes, it is not possible to make firm conclusions. But this is a pilot study, and first attempt at a quantitative investigation of simulator validity. We have been

successful in collecting OR data and making effective comparisons to both VR and physical simulators for surgical residents and experts. Our experimental tool and quantitative analysis system is a novel and unique method to assess surgical performance in various environments.

Chapter 5

Conclusions and Recommendations

5.1 Introduction

The goals of the research presented in this document include a quantitative evaluation of the validity of two types of laparoscopic surgical simulators. And to do this, we developed an experimental tool to collect data in the human operating room, and developed a method to fuse the collected kinematics data. A standard laparoscopic surgical tool was modified, and a bracket designed to accommodate the various sensors used for data collection to collect surgeon motor behaviour in the operating room. Over a period of five months, performance measure data was collected from the operating room, virtual reality simulator and physical simulator for three surgical residents. This data was compared within and between subjects and contexts. By comparing the simulator behaviour to the OR behaviour, we were able to investigate the performance validity of the simulators. This surgical resident data was then compared to expert surgeon behaviour as analysed by Kinnaird (2004). This comparison aided in the evaluation of construct validity of the two simulators. The overall system was initially developed by McBeth (2002), improved upon by our group (Brouwer 2004, Kinnaird 2004), and will be furthered by Sayra M. Cristancho to achieve our overall goal of creating a surgical performance measure database.

5.2 Review of Research

The following sections review and summarize the research conducted in this project. We have covered many areas of study, and will present each in a summarized section.

5.2.1 Experimental Surgical Tool

The design and development of an experimental surgical tool was completed in partnership with Catherine Kinnaird, to create a total system capable of measuring and collecting high frequency continuous kinematics and force/torques of the surgical tool tip. From literature searches, it is thought this was the first time that this variety of sensors was attached to a surgical tool for use in the human operating room.

There were a few different criteria for the hybrid surgical tool to be created. The incorporation of delicate sensors, and the acceptance of the tool for use in the OR by surgeons and the OR staff were of utmost importance. The biggest challenge was to be able to mount the force/torque (F/T) sensor onto the surgical tool shaft. With the aid of volunteer Brandon Lee (engineering graduate), we were able to create a bracket to mount the F/T sensor off-axis and still be able to transmit forces through the sensor without changing the function of the laparoscopic tool. This bracket also allowed for the mounting of the kinematics sensors. The custom-designed experimental tool allowed for high frequency continuous data collection of kinematics and F/T measures.

The kinematics portion of the system consists of both optoelectronic and electromagnetic position tracking systems. These two data streams are collected separately with their respective tracking systems and software. Another objective of this project was to be able to combine these two data sets into one continuous high frequency stream. In this fashion, we can take advantage of the accuracy of the optical system and the high frequency continuity of the magnetic system. This fusion of the datasets is a simple yet efficient method to obtain accurate continuous high frequency kinematics performance measures. It is also a large improvement over the previous kinematics data collection system previously used in our lab.

The force/torque system is the newest part of the total data collection system. This component was incorporated into the quantitative performance measure system, and will need some improvement in future studies. Issues with friction in the tool shaft and bracket design problems have led possibly to misleadingly high force data. A redesign of the bracket and possibly a new F/T sensor that can be mounted on the tool shaft would help in the problems that we dealt with. There is also the issue of trocar interaction forces that was not included in this study. These interaction forces could contribute significantly to the force values that we have measured.

Another issue that caused problems with our data collection and processing system was the use of electrocautery during the surgical procedures. The surgeons commonly use cautery to cut and coagulate tissues to minimize the amount of blood in the surgical field. But our sensors were affected by this high frequency high voltage electrocautery current, and would lead to a

lot of noise in the data of both the kinematics (magnetic sensor) and F/T (strain gauges). We developed post-processing techniques to remove these noisy portions of data, and also manually removed some parts also.

5.2.2 Data Collection

5.2.2.1 The Operating Room

Attempting to collect data during a live human operation is a difficult undertaking that is fraught with logistical nightmares: equipment failure, patient consent, surgeon scheduling, hospital strike, and other numerous problems that seemed to crop up weekly. The original plan was to collect data at least twice per week. But instead, we were only able to collect data once every few weeks. Due to these problems, this was the main reason on why we had to switch our focus from a transfer of training study to a validity study.

The created data collection and analysis system is a good start into the realm of surgeon motor behaviour analysis and measurement. But in its present state, it is not feasible to collect data often or to process a large amount of data in a reasonable amount of time. An average of 15 hours minimum was required to process the acquired 15-30minutes of OR data into a usable form. Although we tried to minimize the disturbance in the OR, our large amount of equipment, and the two researchers required to operate the system, did receive complaints from the OR staff. The actual size of the OR is relatively small, and by adding the extra equipment and people, we were sometimes “in the way”, and created a hassle for the staff. We had also planned to do calibrations immediately post-operatively, but due to logistics, this was not always possible.

5.2.2.2 The Experimental Surgical Tool

Our custom-designed experimental surgical tool is one of the first such tools to be used in a human operating room to monitor surgeon motor behaviour during a laparoscopic cholecystectomy. This tool was used to collect data successfully in the OR a total of three times, although four trials were attempted. This tool was designed in consultation with expert surgeons, and was designed with the ease of use for the surgeon in mind. So although we tried to meet the criteria set by the surgeons, the end result was an “awkward” tool as commented by all the surgeons, expert and surgical residents. The main concern was the size and weight of the

mounting bracket. The bracket was designed to be as lightweight as possible but due to the placement of the sensors, it tended to be weighted significantly on one side, impeding the normal roll direction around the tool shaft. Also because of the wires coming from the multiple sensors, they also tended to keep the tool from rolling around, and always swinging back to the original position. This experimental tool is a very good first step in the creation of an instrumented surgical tool capable of collecting kinematics and F/T measurements in a human OR.

5.2.2.3 Simulators

The physical simulator data collection process utilized the same system as for the operating room data collection but without the same logistical problems. The data was easier to post-process, as there were not issues of electrocautery noise. The physical simulator consisted of the dissection of a mandarin orange using standard laparoscopic setup (tower and camera), and was conducted in the Centre of Excellence for Surgical Education and Innovation (CESEI) at Vancouver General Hospital (VGH).

The virtual reality (VR) simulator data collection process was comparatively simple, although it is noted that Iman Brouwer spent a lot of time configuring and calibrating this simulator. The continuous high frequency kinematics and force/torque data is directly extracted from and formatted by the VR simulator software. This data was also collected with the aid of Iman Brouwer in CESEI at VGH.

5.2.3 Data Fusion

One of the objectives of this project was to create a high frequency continuous data stream of kinematics data. We are able to achieve this goal by taking the data gathered from our two position sensors, and fusing them into one data set. So after the data is gathered from the operating room or physical simulator contexts, registered and time synchronized, the fusion process is started. It is a simple, yet effective method. By using the advantages of both systems, we are able to create a data set that is accurate, high frequency and continuous. We have found a large decrease in error over the previously implemented interpolation technique.

5.2.4 Performance Measures

The quantitative measurement of surgeon performance is of utmost importance to both the public and surgical community. It is necessary to know how our surgeons are performing, and not the simple fact that they can do these procedures. Some of the quantitative measures used to assess surgical performance include completion time, force/torques, kinematics, and ergonomics (Chung 1998, Hanna 1998, McBeth 2002, Rosen 2001, Sackier 1998). Our system to capture kinematics and force/torque data *in vivo* is very innovative. There were twenty-six performance measures that were investigated: velocity, acceleration, jerk, distance from mean (D mean), and force in the following tool tip directions (axial, grasp, translation, transverse, absolute, roll). The performance measures that we have chosen seem to be reliable as there was little variability between surgical residents in the same environment. This further supports the data found by Kinnaird (2004).

We had a total of 26 performance measures to analyze. It is possible that we may not need this wide of a selection of measures, as we were able to make generalizations by looking at the force, \bar{d} , and kinematics measures. Just looking at the velocity, \bar{d} and force measures may give us enough detail to conduct comparisons. Also we chose to look at five tool tip directions for these measures. This also may not be necessary, and we could choose to just analyze one tip direction. The force measures were the only one that showed differences in all five tool tip directions. Perhaps in the future, the measures could be reduced to as shown in Figure 5.1.

	Axial (z)	Grasp (y)	Translate (x)	Transverse	Absolute	Rotation
D mean	☆				☆	☆
Velocity	☆				☆	☆
Force	☆	☆	☆	☆	☆	☆

Figure 5.1: *New performance measures. We may be able to reduce the number of performance measures. This will decrease post-processing time, and make comparisons easier and more generalized.*

5.2.5 Context Comparisons

Comparisons were made over the three settings and amongst the subjects. These comparisons helped to establish our construct and performance validity assessments. The Kolmogorov-Smirnov (KS) statistic was used to calculate the differences between these contexts. By looking at these D-values from the KS statistic, we can quantitatively compare the differences without making any assumptions about the distribution of the data. We collected data from surgical residents, and used the expert data collected by Kinnaird (2004) for our various comparisons.

The comparisons that were made led us to the following conclusions and new ideas:

- OR intraprocedural context can make a difference but is not consistent between subjects in which segments are similar (i.e., segment 1 not always the most different than segment 2 and segment 3 as was found by Kinnaird (2004))
 - Each segment has a different goal in mind
 - Other variables could affect (e.g., patient anatomy, complications)
- Residents show very low intertrial variability in the VR simulator
 - VR simulator is very repeatable and structured environment
 - Each test is the same as previous, and residents complete task in similar fashion
- Intersubject intrasetting comparisons show increasing differences from VR simulator to physical simulator to OR (most repeatable to least repeatable environments)
- Intersetting comparisons show that VR simulator is the most different from the OR and physical simulator contexts, and the physical simulator is relatively similar to the OR. Physical simulator shows fair performance validity.
- Interlevel differences are seen leading to a suggestion of fair construct validity for the physical simulator
- VR simulator differences were very low between skill levels, so does not show construct validity
- Performance measures of force and distance from mean (\bar{d}) show the most sensitivity in context comparisons

5.2.6 Simulator Validation

A valid simulator is one that correctly represents the setting in which it is trying to emulate. By making comparisons between the OR and the two simulated settings (VR and physical), we can see if either of the simulators does a reasonable job of re-creating OR kinematics and forces.

When making interesting comparisons between the VR and physical simulator and investigating performance validity, we found that the VR simulator was the most different from the OR context. The residents treated the physical simulator more similar to the OR. This is an important note, as Kinnaird (2004) found that the expert surgeons treated the two simulators about equally different from the OR context. In this project, we can suggest that the residents treat the physical simulator relatively the same as the OR, and this simulator does show fair performance validity. For the surgical education program, this is a significant find as the residents are practicing similar kinematics in the physical simulator as in the OR. Another important factor here is the cost of each simulator. The VR simulator is ~\$50 000 while the physical simulator is ~\$1. We did find the residents tend to move slower and use a lot less force in the VR simulator as compared to the other two contexts.

One objective was to investigate the construct validity of both the physical and VR simulators. The method we have chosen is to study the differences between expert surgeons and surgical residents in both these environments. If a simulator can detect differences between skill levels, it is considered to show construct validity. In our interlevel comparisons, we see some differences in the physical simulator. The VR simulator shows very little interlevel differences. It is interesting to note that the VR simulator shows small differences between the two skill levels ($D < 0.3$) in the kinematics measures, so could almost be considered "similar" between the residents and experts. Even though both simulators received mixed reviews from the residents and experts, according to our data, the physical simulator would be a better context for training, as it does show fair construct validity.

5.3 Recommendations

Recommendations and improvements were suggested in consultation with fellow researchers, surgeons, and operating room staff. Although we have created a good first approach at

gathering and analyzing surgeon motor behaviour in the operating room, future studies could be even further improved with some modifications.

5.3.1 Software

- “One button” operation for collection of all data from all sensors from one laptop computer. This was the original plan, but due to inherent multiple serial port issues with Matlab, this was not possible.
- Custom designed software to automate the post-processing (data registration and calibration). This was a laborious and tedious task for each set of OR data. This minimum 15-hour task could be shortened into a more reasonable timeframe.
- Automatic data synchronization programs are commercially available to remove the human error aspect of data synching manually and visually.
- Custom-designed software to automatically recognize when electrocautery is used during surgery (either during or post-process), and to compensate for these parts of the data stream, but either filtering or automatic removal of the noisy data.

5.3.2 Hardware

- Sensor bracket redesign by making it more compact and allowing for complete normal use of the surgical tool, especially in the roll around the tool axis direction. Create out of a more rigid non-conductive material to prevent wear and allow for better force transmission to the F/T sensor.
- Wireless sensors would be most optimal, as this would remove the issue of having many wires hanging down and affecting the weighting and turning of the surgical tool.
- Improvements to the strain gauge system, as it seemed to be most affected by the electrocautery during the surgical procedure. A different configuration or instrumentation amplifier may be able to solve these problems.
- A variety of surgical tool tips could be purchased. This would allow for more tasks of a procedure to be analyzed, and not only the dissection. The electrocautery hook or spatula would be the next most used tool during a laparoscopic cholecystectomy.

5.3.3 OR Data Collection

- Keep a dedicated OR cart with all data collection equipment, and a stand for the Polaris camera, and the video camcorder.
- Minimize disturbance to the OR staff by only having one researcher in the room for the entire procedure. The 2nd researcher should leave the room once the computer system is up and running.
- Arrive early for all OR trials to double and triple check all equipment is functioning.
- Always book for the first operation in the morning in the largest OR. This allows for more time to setup and check equipment.
- Allowance for intraoperative dynamic tracking of the magnetic transmitter. This would allow for optimal placement of the transmitter. A passive optical marker attached to the transmitter may be useful.

5.3.4 Simulators

- Physical simulator improvements would be similar to the software and hardware recommendations. Also, choosing the proper mandarin orange for the dissection task is also important. The orange must not be too juicy or firm, as this makes for a difficult and messy dissection task. Another improvement would be to create a permanent mounting surface to place the mandarin orange.
- Virtual reality simulator improvements would need to be discussed with the commercial manufacturers. One immediate modification is to improve the force feedback effects.

5.3.5 Other Recommendations

- Account for F/T trocar interaction forces on the surgical tool shaft.
- Automated performance measure extraction.

5.4 Partner & Future Studies

Our new experimental surgical tool and analysis system is a worthy first contribution into the study of surgeon motor behaviour. The methods we have created are feasible, and with some improvements, a better system could be created. The next step would be to implement as many of the recommendations as possible, and to collect more OR data.

Some areas of immediate study that could be investigated with an improved system include:

- Which performance measures are the most sensitive, valid and reliable to assess surgeons?
- Does training in the virtual reality and/or physical simulator lead to improved performance in the operating room? (This was our original objective question, and still needs to be addressed).

The longer-term goal of the research in our lab (Neuromotor Control Laboratory, University of British Columbia, Canada) is to eventually create a surgical skills database. This would involve collecting performance measures from many different surgical skill levels ranging from the very novice to the expert surgeon. As this database increases in size, a surgeon from any skill level could see how they compared to others of their own skill level. For example, a PGY2 resident could see how they compared to others of the same year level. This could be done for both operating room performance measures as well as in the simulators. A surgeon could also see the specific areas in which they need to improve or where they excel as compared to others.

In conjunction with this current research project, there are two projects within our lab occurring also studying different aspects of surgical simulators. These two projects and the one described in this thesis all fit together (Figure 5.2) to set-up the framework to eventually create the surgical skills databases as mentioned above. Catherine Kinnaird began the investigation of the validity of both VR and physical simulators with expert surgeon subjects (Kinnaird 2004). Iman Brouwer studied the minimum technological requirements for a virtual reality simulator, and how haptic quality affects simulator performance (Brouwer 2004).

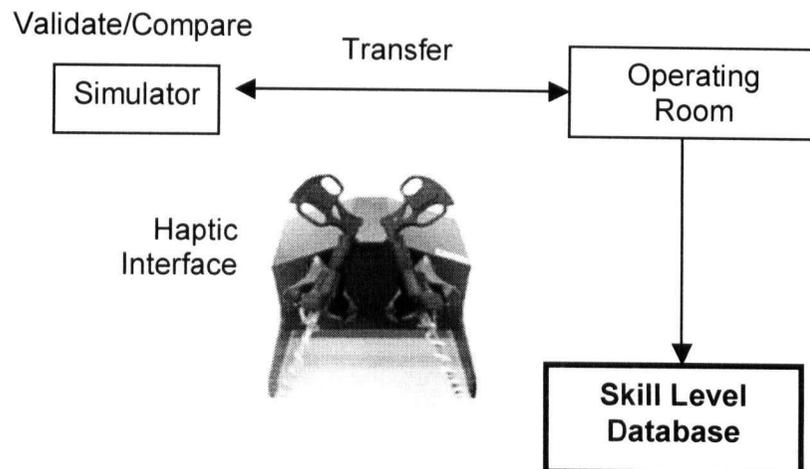


Figure 5.2: *Concurrent research projects at the Neuromotor Control Laboratory (University of British Columbia). The studies are: 1) Transfer of training from simulator to operating room, 2) Validation of physical and virtual reality simulators, 3) Minimum technological requirements for a virtual reality simulator, 4) Skill Level Database*

In a larger more elaborate study, Sayra M. Cristancho will be continuing our projects by using the experimental tool and data collection system to assess more surgeons in the OR. This will help in our final global goal in the creation of the surgical skill level database. Currently, our data analysis methods are time consuming, so Ms. Cristancho will be working on automating the analysis process to be able to collect and process more data in a reasonable amount of time. Many more procedures could be analyzed, and a better picture of surgeon motor behaviour can be obtained.

Using the data collected in this study and that of Catherine Kinnaird, surgical resident Dr. Hamish Hwang will do an analysis of surgical error during a laparoscopic cholecystectomy. He is analyzing the laparoscopic video and the tool tip kinematics and forces/torques data to draw conclusions about qualitative video and quantitative performance measures as they relate to surgical errors.

Another surgical resident, Dr. Hanna Piper, will be looking at the feasibility of using our hybrid experimental tool to analyze the general surgery curriculum new training modules. They

will be using our physical simulator mandarin orange model as one of the surgical educational modules for the University of British Columbia surgical resident training program.

As is seen with our partner and future studies, our project goals have been met and exceeded. The results presented here and in our partner studies provide a significant contribution into the realm of surgical simulator assessment and education. The foundation for a method and system to collect quantitative human operating room performance measures, and to assess construct and performance validity of laparoscopic surgical simulators has been created and used with success.

List of Terms

Abs	Absolute tool tip direction
Accel'	acceleration
BCIT	British Columbia Institute for Technology
β_i	block of length l of dependent data
β_i^*	block of length l randomly resampled from original block set
$\{\beta_1, \dots, \beta_N\}$	blocks of dependent data created from original dependent dataset
$\{\beta_1^*, \dots, \beta_k^*\}$	resampled blocks of dependent data from original block set
CA	cystic artery
CESEI	Center of Excellence for Surgical Education and Innovation
CPD	cumulative probability distribution
CPD(D_{RS-ref})	cumulative probability distribution for the bootstrapped data resampled from itself and compared to the reference
CD	cystic duct
CDD	cystic duct dissection
D	Kolmogorov-Smirnov statistic difference measures
D_{cr}	critical D-measure at the 95 th percentile if CPD(D_{RS-ref})
D_{1-2}	Kolmogorov-Smirnov statistic between two CPD's
D mean	distance from mean
ESU	electrosurgical unit
Ex 1	expert surgeon 1
Ex 2	expert surgeon 2
F/T	force/torque
GB	gallbladder
GBD	gallbladder dissection
GCV	Generalized Cross Validation
GUI	graphical user interface
Hz	Hertz
k	length of resampled block set – $k=N/l$
KS	Kolmogorov-Smirnov
l	length of an individual block for MBB
MBB	Moving Block Bootstrap
mag	magnetic data
mm	millimeters
MDMArray	Multi Dimensional Marker Array

N	Newtons
N	length of block set – by default $N=n-1+1$
NCL	Neuromotor Control Laboratory
n	length of original data set
Nm	Newton meters
OR	operating room
opt	optical data
PC	personal computer
PGY	post-graduate year (for surgical residents)
Phy	physical simulator
RF	radio frequency
RMS	root mean square (error)
rad	radians
r	number of bootstrapping cycles
Res 1	surgical resident 1
Res 2	surgical resident 2
Res 3	surgical resident 3
Roll	rotation (about the experimental tool axis – see Rot)
s	seconds
synch	synchronization
Trans	Transverse tool tip direction
V	Volts
vel	velocity
VR	virtual reality simulator
x	Translation direction of tool tip frame
X_i	data at point i
X_i^*	resampled data at point i
$\{X_1, \dots, X_n\}$	original data set
$\{X_1^*, \dots, X_n^*\}$	resampled data set – created from the resampled block set
UBC	University of British Columbia
y	Grasping direction of tool tip frame
z	Axial direction of tool tip frame
3D	three-dimensional

Bibliography

Adrales, G. L., Chu, U. B., Witzke, D. B., Donnelly, M. B., Hoskins, D., Mastrangelo, M. J., Jr., Gandsas, A., Park, A. E. (2003). Evaluating minimally invasive surgery training using low-cost mechanical simulations. *Surgical Endoscopy* **17**, 580-585.

Adrales, G.L., Park, A.E., Chu, U.B., Witzke, D.B., Donnelly, M.B., Hoskins, J.D., Mastrangelo, M.J. Jr, Gandsas, A. (2003). A valid method of laparoscopic simulation training and competence assessment. *The Journal of Surgical Research* **114**, 156-162.

Ahlberg, G., Heikkinen, T., Iselius, L., Leijonmarck, C. E., Rutqvist, J., Arvidsson, D. (2002). Does training in a virtual reality simulator improve surgical performance? *Surgical Endoscopy* **16**, 126-129.

Anastakis, D. J., Regehr, G., Reznick, R. K., Cusimano, M., Murnaghan, J., Brown, M. Hutchison, C. (1999). Assessment of technical skills transfer from the bench training model to the human model. *American Journal of Surgery* **177**, 167-170.

Atsma, W. ATI Force Transducer driver. http://www.mech.ubc.ca/~wastma/ATI-FT_driver/
Last accessed April 12, 2004.

Auffrey, A.L., Mirabella, A. Siebold, G.L.(2001). Transfer of Training Revisited, Advanced Training methods Research Unit, U.S Army Research Institute for the Behavioural and social Sciences, ARI Research Note 2001-10, July 2001.

Ballantyne, G.H. (2002). The pitfalls of laparoscopic surgery: challenges for robotics and telerobotic surgery. *Surgical laparoscopy, endoscopy & percutaneous techniques* **12**, 1-5.

Bann, S., Datta, V., Khan, M., Darzi, A. (2003). The surgical error examination is a novel method for objective technical knowledge assessment. *American Journal of Surgery* **185**, 507-511.

Berguer, R., Forkey, D.L., Smith, W.D. (1999). Ergonomic problems associated with laparoscopic surgery. *Surgical Endoscopy* **13**, 466-468.

Birkfellner, W., Watzinger, F., Wanschitz, F., Ewers, R., Bergmann, H. (1998). Calibration of tracking systems in a surgical environment. *IEEE Transactions on Medical Imaging* **17**, 737-742.

Blaiwes, A. S. (1984). Training Effectiveness Evaluation and Utilization Demonstration of a Low Cost Cockpit Procedures Trainer (Report No. NAVTRAEQUIPCEN 78-C-001301). Pensacola, Fla.: Seville Training Systems.

Bloom, M.B., Rawn, C.L., Salzberg, A.D., Krummel, T.M. (2003). Virtual reality applied to procedural testing: the next era. *Annals of Surgery* **237**, 442-448.

Bridges, M., Diamond, D. L. (1999). The financial impact of teaching surgical residents in the operating room. *American Journal of Surgery* **177**, 28-32.

Brouwer, I. (2004). Cost-performance trade-offs in haptic hardware design. MSc Thesis, University of British Columbia, Vancouver, BC, Canada.

Cao, C. G., MacKenzie, C. L., Ibbotson, J. A., Turner, L. J., Blair, N. P., Nagy, A. G. (1999). Hierarchical decomposition of laparoscopic procedures. *Studies in Health Technology and Informatics* **62**, 83-89.

Challa, S., Koks, D. (2004). Bayesian and Dempster-Shafer Fusion. *Sadhana* **29**, 145-174.

Chan, A.C., Chung, S.C., Yim, A.P., Lau, J.Y., Ng, E.K., Li, A.K. (1997). Comparison of two-dimensional vs. three-dimensional camera systems in laparoscopic surgery. *Surgical Endoscopy* **11**, 438-440.

Cohen, R., Reznick, R.K., Taylor, B.R., Provan, J., Rothman, A. (1990). Reliability and validity of the objective structured clinical examination in assessing surgical residents. *American Journal of Surgery* **160**, 302-305.

Coue, C., Fraichard, T., Bessiere, P. Mazer, E. (2003). Using Bayesian programming for multi-sensor multi-target tracking in automotive applications. *Int'l Conference on Robotics and Automation*. Taipei, Taiwan. May 12-17, 2003.

Dakin, G.F., Gagner, M. (2003). Comparison of laparoscopic skills performance between standard instruments and two surgical robotic systems. *Surgical Endoscopy* **17**, 574-579.

Datta, V., Chang, A. Mackay, S. Darzi, A. (2002). The relationship between motion analysis and surgical technical assessments. *American Journal of Surgery* **184**, 70-73.

Derossis, A.M., Fried, G.M., Abrahamowicz, M., Sigman, H.H., Barkun, J.S. & Meakins, J.L. (1998). Development of a model for training and evaluation of laparoscopic skills. *American Journal of Surgery*. **175**, 482-487.

Derossis AM, Antoniuk M, Fried GM. (1999). Evaluation of laparoscopic skills: a 2-year follow-up during residency training. *Canadian Journal of Surgery* **42**, 293-296.

Derossis AM, Bothwell J, Sigman HH, Fried GM.(1998). The effect of practice on performance in a laparoscopic simulator. *Surgical Endoscopy* **12**, 1117-1120.

de Visser, H., Heijnsdijk, E.A., Herder, J.L., Pistecky, P.V. (2002). Forces and displacements in colon surgery. *Surgical Endoscopy* **16**, 1426-1430.

Efron, B., Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**: 54-77.

Elmenreich, W. (2002). An introduction to sensor fusion: research report. <http://www.vmars.tuwien.ac.at/frame-papers.html> Last accessed March 20, 2004.

Emam, T.A., Frank, T.G., Hanna, G.B., Cuschieri, A. (2001) Influence of handle design on the surgeon's upper limb movements, muscle recruitment, and fatigue during endoscopic suturing. *Surgical Endoscopy* **15**, 667-672.

Eubanks, T.R., Clements, R. H., Pohl, D., Williams, N., Schaad, D.C., Horgan, S. Pellegrini, C. (1999). An objective scoring system for laparoscopic cholecystectomy. *Journal of the American College of Surgeons* **189**, 566-574.

Faulkner, H., Regehr, G., Martin, J., Reznick, R. (1996). Validation of an objective structured assessment of technical skill for surgical residents. *Academic Medicine: journal of the Association of American Medical Colleges* **71**, 1363-1365.

Feldman, L.S., Hagarty, S.E., Ghitulescu, G., Stanbridge, D., Fried, G.M. (2004). Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *Journal of the American College of Surgeons* **198**, 105-110.

Figert, P. L., Park, A. E., Witzke, D. B., Schwartz, R. W. (2001). Transfer of training in acquiring laparoscopic skills. *Journal of the American College of Surgeons* **193**, 533-537.

Flexman, R. E., Roscoe, S. N., Williams, A. C., Jr., & Williges, B. H. (1972). Studies in pilot training: The anatomy of transfer. *Aviation Research Monographs*, 2(1). Champaign, IL: University of Illinois, Aviation Research Laboratory.

Ford, J.K., Weissbein, D. A. (1997). Transfer of Training: An Updated Review and Analysis, *Performance Improvement Quarterly* **10**, 22-41.

Foxon, M. (1993). A process approach to the transfer of training: The impact of motivation and supervisor support on transfer maintenance. *Australian Journal of Educational Technology* **9**, 130-143

Francis, N.K., Hanna, G.B., Cuschieri, A. (2002). The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester: contrast validity study. *Archives of Surgery* **137**, 841-844.

Francoeur, J.R., Wiseman, K., Buczkowski, A.K., Chung, S.W., Scudamore, C.H. (2003). Surgeons' anonymous response after bile duct injury during cholecystectomy. *American Journal of Surgery* **185**, 468-475.

Frantz, D.D., Wiles, A.D., Leis, S.E., Kirsch, S.R. (2003). Accuracy assessment protocols for electromagnetic tracking systems. *Physics in Medicine and Biology* **48**, 2241-2251.

Fried, G. M., Derossis, A. M., Bothwell, J., Sigman, H. H. (1999). Comparison of laparoscopic performance in vivo with performance measured in a laparoscopic simulator. *Surgical Endoscopy* **13**, 1077-1081, discussion 1082.

Fried, G.M., Feldman, L.S., Vassiliou, M.C., Fraser, S.A., Stanbridge, D., Ghitulescu, G., Andrew, C.G. (2004). Proving the Value of Simulation in Laparoscopic Surgery. *Annals of Surgery* **240**, 518-528.

Gallagher, A. G., Richie, K., McClure, N., McGuigan, J. (2001). Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World Journal of Surgery* **25**, 1478-1483.

Gallagher A.G., Satava, R.M. (2002). Virtual reality as a metric for the assessment of laparoscopic psychomotor skills; learning curves and reliability measures. *Surgical Endoscopy* **16**, 1746-1752

Gallagher, A.G., Richie, K., McClure, N., McCuigan, J.(2001). Objective psychomotor skills assessment of experienced, junior and novice laparoscopists with virtual reality. *World Journal of Surgery* **25**, 1478-1483.

Gallagher, A.G., Lederman, A.B., McGlade, K., Satava, R.M., Smith, C.D. (2004). Discriminative validity of the Minimally Invasive Surgical Trainer in Virtual Reality (MIST-VR) using criteria levels based on expert performance. *Surgical Endoscopy*, **18**, 660-665.

Glinatsis, M.T., Griffith, J.P., McMahon, M.J. (1992). Open versus laparoscopic cholecystectomy: a retrospective comparative study. *Journal of Laparoendoscopic Surgery* **2**, 81-86.

Goff, B.A., Lentz, G.M., Lee, D., Fenner, D., Morris, J., Mandel, L.S. (2001). Development of a bench station objective structured assessment of technical skills. *Obstetrics and Gynecology* **98**, 412-416.

Grantcharov, T. P., Bardram, L., Funch-Jensen, P., Rosenberg, J. Assessment of technical surgical skills. (2002). *European Journal of Surgery* **168**, 139-144.

Grantcharov, T.P., Bardram, L., Funch-Jensen, P., Rosenberg, J. (2003). Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills. *American Journal of Surgery* **185**, 146-149.

Grantcharov, T.P., Kristiansen, V.B., Bendix, J., Bardam, L., Rosenberg, J., Funch-Jensen, P. (2004). Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *British Journal of Surgery* **91**, 46-150.

Gustafsson, F. (2003) <http://www.control.isy.liu.se/~fredrik/isis/positioning.html> Last accessed: August 21, 2004.

Hall, P., Horowitz, J.L., Jing, B.Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika* **82**, 561-574.

Haluck, R.S., Marshall, R.L., Krummel, T.M., Melkonian, M.G. (2001) Are surgery training programs ready for virtual reality? A survey of program directors in general surgery. *Journal of the American College of Surgeons* **193**, 660-5.

Hamilton, E. C., Scott, D. J., Fleming, J. B., Rege, R. V., Laycock, R., Bergen, P. C., Tesfay, S. T., Jones, D. B.(2002). Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills. *Surgical Endoscopy* **16**, 406-411.

Hanna, G.B., Drew, T., Clinch, P., Shimi, S., Dunkley, P., Hau, C., Cuschieri, A. (1997). Psychomotor skills for endoscopic manipulations: differing abilities between right and left-handed individuals. *Annals of Surgery* **225**, 333-338.

Hanna, G.B., Shimi, S.M., Cuschieri, A. (1998), Randomised study of influence of two-dimensional versus three-dimensional imaging on performance of laparoscopic cholecystectomy. *Lancet* **51**, 248-251.

Hannaford, B. (2004). Private Discussion September 2004 at the University of British Columbia, Vancouver, BC, Canada.

Harms, J., Feussner, H., Baumgartner, M., Schneider, A., Donhauser, M., Wessels, G. (2001). Three-dimensional navigated laparoscopic ultrasonography: first experiences with a new minimally invasive diagnostic device. *Surgical Endoscopy* **15**, 1459-1462.

Herline, A., Stefansic, J.D., Debelak, J., Galloway, R.L., Chapman, W.C. (2000). Technical advances toward interactive image-guided laparoscopic surgery. *Surgical Endoscopy* **14**, 675-679.

Hodgson, A.J., Person, J.G., Salcudean, S.E., Nagy, A.G. (1999). The effects of physical constraints in laparoscopic surgery. *Medical Image Analysis* **3** 275-83.

Hubens, G., Coveliers, H., Balliu, L., Ruppert, M., Vaneerdeweg, W. (2003). A performance study comparing manual and robotically assisted laparoscopic surgery using the da Vinci system. *Surgical Endoscopy* **17**, 1595-1599.

Huntsville Gastroenterology Associates (2002). (http://www.huntsville-gastroenterology.com/laparoscopic_cholecystectomy.shtml, Last accessed January 12, 2004.

Hyltander, A., Liljegren, E., Rhodin, P. H., Lonroth, H. (2002). The transfer of basic skills learned in a laparoscopic simulator to the operating room. *Surgical Endoscopy* **16**, 1324-1328.

Johnson, J.L., Schamschula M.P., Inguva, R., Caulfield, H.J., (1998). Pulse-coupled neural network sensor fusion. *Proceedings of SPIE* **3376**, 219-226.

Joice, P. Hanna, G. B. Cuschieri, A. (1998). Errors enacted during endoscopic surgery--a human reliability analysis. *Applied Ergonomics* **29**, 409-414.

Jones, D.B., Brewer, J.D., Soper, N.J. (1996). The influence of three-dimensional video systems on laparoscopic task performance. *Surgical Laparoscopy & Endoscopy* **6**, 191-197.

Jordan, J.A., Gallagher, A.G., McGuigan, J., McClure, N. (2001). Virtual reality training leads to faster adaptation to the novel psychomotor restrictions encountered by laparoscopic surgeons. *Surgical Endoscopy* **15**, 1080-1084.

Kinnaird, C. (2004). A Multifaceted Quantitative Validity Assessment of Laparoscopic Surgical Simulators. MASc Thesis, University of British Columbia, Vancouver, BC, Canada.

- Kopta, J. A. (1971). An approach to the evaluation of operative skills. *Surgery* **70**, 297-303.
- Kunsch, S. N. (1991). Second order optimality of stationary bootstrap. *Statistics and Probability Letters* **11**, 335-341.
- Lahari, S.N. (2003). Resampling methods for dependent data. Springer-Verlag, New York.
- Liu, R.Y., Singh, K. (1992). Moving block, jackknife, and bootstrap capture weak dependence. *Exploring the Limits of the Bootstrap*. Wiley, New York: 225: 248.
- Lujan, J.A., Parrilla, P., Robles, R., Marin, P., Torralba, J.A., Garcia-Ayllon, J. (1998). Laparoscopic cholecystectomy vs open cholecystectomy in the treatment of acute cholecystitis: a prospective study. *Archives of Surgery* **133**, 173-175.
- Marescaux, J., Smith, M.K., Folscher, D., Jamali, F., Malassagne, B., Leroy, J. (2001). Telerobotic laparoscopic cholecystectomy: initial clinical experience with 25 patients. *Annals of Surgery* **234**, 1-7.
- MacRae, H., Regehr, G., Leadbetter, W., Reznick, R.K. (2000). A comprehensive examination for senior surgical residents. *American Journal of Surgery* **179**, 190-193.
- Martin, J.A., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery* **84**, 273-278.
- Martin, M., Scalabrini, B., Rioux, A., Xhignesse, M.A. (2003). Training fourth-year medical students in critical invasive skills improves subsequent patient safety. *The American Surgeon* **69**, 437-440.
- McBeth, P.B., (2002). A Methodology for Quantitative Performance Evaluation in Minimally Invasive Surgery. MAsc Thesis, University of British Columbia, Vancouver, BC, Canada.
- McCarthy, A., Harley, P. Smallwood, R. (1999). Virtual arthroscopy training: do the "virtual skills" developed match the real skills required? *Studies in Health Technology and Informatics* **62**, 221-227.
- McDougall, E.M., Soble, J.J., Wolf, J.S. Jr., Nakada, S.Y., Elashry, O.M., Clayman, R.V. (1996). Comparison of three-dimensional and two-dimensional laparoscopic video systems. *Journal of Endourology* **10**, 371-374.
- McNatt, S.S., Smith, C.D.(2001). A computer-based laparoscopic skills assessment device differentiates experienced from novice laparoscopic surgeons. *Surgical Endoscopy* **15**, 1085-1089.
- Milne, A.D., Chess, D.G., Johnson, J.A., King, G. J.W. (1996). Accuracy of an electromagnetic tracking device: a study of optimal operating range and metal interference. *Journal of Biomechanics* **29**, 791-3.

Mishra, R.K., http://www.laparoscopyhospital.com/history_of_laparoscopy.htm Last accessed January 12, 2004.

Moore, K. 2002 <http://www.bleep.demon.co.uk/SimHist1.html> Last accessed January 15/04

Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery, *BMJ* 327, November 2003, 1032-7.

Morimoto, A.K., Foral, R.D., Kuhlman, J.L., Zucker, K.A., Curet, M.J., Bocklage, T., MacFarlane, T.I., Kory, L. (1997). Force sensor for laparoscopic Babcock. *Studies in Health and Technology Informatics* 39, 354-61.

Nagy, A.G., Poulin, E.C., Girotti, M.J., Litwin, D.E., Mamazza, J. (1992). History of laparoscopic surgery. *Canadian Journal of Surgery* 35, 271-4.

Nakamoto, M., Sato, Y. Tamaki, Y., Nagano, H. Miyamoto, M. Sasama, T. Monden, M. Tamura, S. (2000). Magneto-optic hybrid 3D sensor for surgical navigation. *Lecture Notes in Computer Science* 1935, 839-848.

Nelson, M.S. (1990). Models for teaching emergency medicine skills. (1990). *Annals of Emergency Medicine* 19, 333-335.

O'Toole, R.V., Playter, R.R., Krummel, T.M., Blank, W.C., Cornelius, N.H., Roberts, W.R., Bell, W.J., Raibert, M. (1999). Measuring and developing suturing technique with a virtual reality surgical simulator. *Journal of the American College of Surgeons* 189, 114-127.

Paisley, A.M., Baldwin, P.J., Paterson-Brown, S. (2001). Validity of surgical simulation for the assessment of operative skill. *British Journal of Surgery* 88, 1525-1532.

Perez, A., Zinner, M.J., Ashley, S.W., Brooks, D.C., Whang, E.E. (2003). What is the value of telerobotic technology in gastrointestinal surgery? *Surgical Endoscopy* 17, 811-813.

Périssat, J. (1995). Laparoscopic surgery in gastroenterology: an overview of recent publications. *Surgical Endoscopy* 27, 106-118.

Perkins, N., Starkes, J.L., Lee, T.D., Hutchison, C. (2002). Learning to use minimal access surgical instruments and 2-dimensional remote visual feedback: how difficult is the task for novices? *Advances in health sciences education: theory and practice* 7, 117-131.

Person, J.G. (2000). A Foundation for the Design and Assessment of Improved Instruments for Minimally Invasive Surgery. MAsc Thesis, University of British Columbia, Vancouver, BC, Canada.

Pessaux, P., Regenet, N., Tuech, J.J., Rouge, C., Bergamaschi, R., Arnaud, J.P. (2001). Laparoscopic versus open cholecystectomy: a prospective comparative study in the elderly with acute cholecystitis. *Surgical laparoscopy, endoscopy, & percutaneous techniques* 11, 252-255.

- Poulin, F., Amiot, L. P (2002). Interference during the use of an electromagnetic tracking system under OR conditions. *Journal of Biomechanics* **35**, 733-737.
- Prasad, A., Foley, R. J.(1996). Day care laparoscopic cholecystectomy: a safe and cost effective procedure. *European Journal of Surgery* **162**, 43-46.
- Prystowski, J.B. (1999). A virtual reality simulator for intravenous catheter placement. *American Journal of Surgery* **177**, 171-175.
- Regehr, G., MacRae, H., Reznick, R.K., Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine* **73**, 993-997.
- Reznick, R., Regehr, G., MacRae, H., Martin, J., McCulloch, W. (1997). Testing technical skill via an innovative "bench station" examination. *American Journal of Surgery* **173**, 226-230.
- Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., Penn, P. R., Ambihapahan, N. (2000). Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics* **43**, 494-511.
- Rosen, J., MacFarlane, M., Richards, C., Hannaford, B., Sinanan, M. (1999). Surgeon-tool force/torque signatures--evaluation of surgical skills in minimally invasive surgery. *Studies in Health Technology and Informatics* **62**, 290-296.
- Rosen, J., Hannaford, B., Richards, C.G. Sinanan, M.N. (2001). Markov Modeling of Minimally Invasive Surgery Based on Tool/Tissue Interaction and Force/Torque Signatures for Evaluating Surgical Skill, *IEEE Transactions on Biomedical Engineering* **48(5)**, 579-91.
- Rosen J, Solazzo M, Hannaford B, Sinanan M. (2002). Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. *Computer Aided Surgery* **7**,49-61.
- Rosser, J.C., Wood. M., Payne, J.H., Fullum, T.M., Lisehora, G.B., Rosser, L.E., Barcia, P.J., Savalgi, R.S. (1997). Telementoring. A practical option in surgical training. *Surgical Endoscopy* **11**, 852-5.
- Rosser JC, Rosser LE, Savalgi RS. (1997). Skill acquisition and assessment for laparoscopic surgery. *Archives of Surgery* **132**, 200-4.
- Risucci, D., Cohen, J. A., Garbus, J. E., Goldstein, M., Cohen, M. G. (2001). The effects of practice and instruction on speed and accuracy during resident acquisition of simulated laparoscopic skills. *Current Surgery* **58**, 230-235.
- Ruurda, J.P., Visser, P.L., Broeders, I.A. (2003). Analysis of procedure time in robot-assisted surgery: comparative study in laparoscopic cholecystectomy. *Computer Aided Surgery* **8**, 24-29.
- Ruurda, J.P., Broeders, I.A., Simmermacher, R.P., Rinkes, I.H., Van Vroonhoven, T.J. (2002).

Feasibility of robot-assisted laparoscopic surgery: an evaluation of 35 robot-assisted laparoscopic cholecystectomies. *Surgical Laparoscopy, Endoscopy, and Percutaneous Techniques* **12**, 41-45.

Schijven, M.m Jakimowicz, J. (2002). Face, expert, and referent validity of the Xitact LS500 Laparoscopy Simulator. *Surgical Endoscopy* **16**, 1764-1770.

Schijven, M., Jakimowicz, J. (2003). Construct validity: experts and novices performing on the Xitact LS500 laparoscopy simulator. *Surgical Endoscopy* **17**, 803-810.

Scott, D. J., Bergen, P. C., Rege, R. V., Laycock, R., Tesfay, S. T., Valentine, R. J., Euhus, D. M., Jeyarajah, D. R., Thompson, W. M., Jones, D. B. (2000). Laparoscopic training on bench models: better and more cost effective than operating room experience? *Journal of the American College of Surgeons* **191**, 272-283.

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., Satava, R. M. (2002). Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of Surgery* **236**, 458-463, discussion 463-464.

Smith, S. G., Torkington, J., Brown, T. J., Taffinder, N. J., Darzi, A. (2002). Motion Analysis. *Surgical Endoscopy* **16**, 640-645.

Smith-Jentsch, K. A., Salas, E., Brannick, M. T. To transfer or not to transfer? (2001). Investigating the combined effects of trainee characteristics, team leader support, and team climate. *The Journal of Applied Psychology* **86**, 279-292.

Starkes, J.L., Payk, I., & Hodges, N.J. (1998). Developing a standardized test for the assessment of suturing skill in novice microsurgeons. *Microsurgery* **18**, 19-22.

Stefansic, J.D., Bass, W.A., Hartmann, S.L., Beasley, R.A., Sinha, T.K., Cash, D.M., Herline, A.J., Galloway, R.L. Jr. (2002). Design and implementation of a PC-based image-guided surgical system. *Computer Methods and Programs in Biomedicine* **69**, 211-224.

Strom, P., Kjellin, A., Hedman, L., Johnson, E., Wredmark, T., Fellander-Tsai, L. (2003). Validation and learning in the ProCedicus KSA virtual reality surgical simulator. *Surgical Endoscopy* **17**, 227-231.

Szalay, D., MacRae, H., Regehr, G., Reznick, R. (2000). Using operative outcome to assess technical skill. *American Journal of Surgery* **180**, 234-237.

Taffinder, N., Darzi, A., Smith, S., Taffinder, N. (1999). Assessing operative skill. Needs to become more objective. *BMJ*. **318**. 887-8.

Taffinder, N., Sutton, C., Fishwick, R.J., McManus, I.C., Darzi, A. (1998). Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. *Studies in Health Technology and Informatics* **50**, 124-130.

- Teague, R. C., Gittelman, S. S. Park, O. (1994). A review of the literature on part-task and whole-task training and context dependency (ARI Technical Report 1010) Alexandria ,VA; U.S Army Research Institute for the Behavioural and Social Sciences
- Tracey, M. R., Lathan, C. E. (2001). The interaction of spatial ability and motor learning in the transfer of training from a simulator to a real task. *Studies in Health Technology and Informatics* **81**, 521-527.
- Treat, M. (1996). A surgeon's perspective on the difficulties of laparoscopic surgery. In *Computer-Integrated Surgery* (Taylor, R.H., Lavallée, S., Burdea, G.C., and Mösges, R., eds.), MIT Press, Cambridge, MA., 559-560.
- Verner, L., Oleynikov, D., Holtmann, Haider, Zhukov, L. (2002). Measurements of the Level of Surgical Expertise Using Flight Path Analysis from *da Vinci* Robotic Surgery System. <http://webmedia.unmc.edu/medicine/morien/mis/FlightAnalysis.pdf> Last accessed January 21, 2004.
- Vuilleumier, H., Halkic, N. (2003). Implementation of robotic laparoscopic cholecystectomy in a university hospital. *Swiss medical weekly* **133**, 347-349.
- Way, L.W., Stewart, L., Gantert, W., Liu, K., Lee, C.M., Whang, K., Hunter, J.G. (2003). Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. *Annals of Surgery* **237**, 460-469.
- Wentink M, Stassen LP, Alwayn I, Hosman RJ, Stassen HG. (2003). Rasmussen's model of human behavior in laparoscopy training. *Surgical Endoscopy* **17**,1241-1246.
- Williams, A. C., Jr., & Flexman, R. E. (1949). Evaluation of the school link as an aid in primary flight instruction. University of Illinois Bulletin, 46, (71), (Aeronautics Bulletin: No. 5), University of Illinois.
- Woltring, H. J. (1986). A Fortran Package for Generalized Cross-Validatory spline Smoothing and Differentiation. *Advances in Engineering Software* **8(2)**, 142-151.
- Wu, H., Siegel, M., Stiefelhagen, R. Yang, J., (2002). Sensor Fusion Using Dempster-Shafer Theory. IEEE Instrumentation and Measurement Technology Conference, Anchorage, AK, USA. May 21-23, 2002.
- Yamauchi, Y., Yamashita, J., Morikawa, O., Hashimoto, R., Mochimaru, M., Fukui, Y., Uno, H., Yokoyama, K. (2002). Surgical Skill Evaluation by Force Data for Endoscopic Sinus Surgery Training System. *Lecture Notes in Computer Science* **2488**, 44-51.
- Zeyada Y, Hess RA.(2000). Modeling human pilot cue utilization with applications to simulator fidelity assessment. *Journal of Aircraft* **37**, 588-97.

Appendix A

OR Study Experimental Protocol and Data Acquisition Procedures

A.1 Experimental Protocol

Attending Surgeons:	Dr. Alex Nagy Dr. Neely Panton
Surgical Residents:	Dr. Ed Chang Dr. Naisan Garraway Dr. Kathy Hsu
Researchers:	Joanne Lim Catherine Kinnaird Iman Brouwer (stand-by) Sayra M. Cristancho (stand-by)
Location:	University of British Columbia Hospital
Procedure:	MIS cholecystectomy

Study Protocol:

This is the protocol for the laparoscopic surgery performance evaluation study of the three specified surgical residents performing a MIS cholecystectomy, with an attending surgeon available. Before the patient arrives in the OR, all the equipment is checked and initialized. The resident is asked to scrub and enter the OR while the patient is being anesthetized. One of the researchers also scrubs, in order to affix Opsite™ and Mepore™ to the force/torque sensor mounted on the modified laparoscopic surgical tool. This is to prevent foreign liquids and substances from contaminating the force/torque sensor. The researcher remains scrubbed to be available to make any adjustments to the tool, and to hand off the wires from the surgical tool to the other researcher outside of the sterile field.

When the modified tool is ready to be used in the surgery as noted by the attending surgeon or resident, the researcher outside the sterile field performs a test to ensure the motion capture equipment is functioning. Once the equipment is tested and confirmed operational, the researcher informs the surgeon that they are ready to begin recording. The researcher begins collecting data from the sensors and begins recording the operation using both the external video camera and laparoscopic camera. The equipment records for the entire surgery. When the laparoscopic portion of the surgery is completed, the surgeon informs the researcher, and data collection can be stopped.

As the patient is being sutured, if possible and not too intrusive, the scrubbed researcher holds and manipulates the surgical tool in various positions for synchronization purposes. The surgical tool is held in a horizontal position and a vertical movement is done to strike the tool against a hard surface (i.e. surgical bed). This movement is recorded on both kinematics sensors, and is used for time synchronization. The tool handles are also squeezed while the tool is moved to aid in force synchronization. After these calibrations are completed, all the systems are shut down.

If the surgeon feels uncomfortable using the modified laparoscopic tool at any time during the surgery, they are free to stop the experiment and return to using the traditional non-modified surgical tool. Approval for this experiment was granted through the University of British Columbia Clinical Research Ethics Board and the University of British Columbia Hospital.

The Sterile Supply Department and the Biomedical Engineering Department had approved all equipment and instrumentation.

A.2 Equipment List

The following is the list of hardware and software components required for the OR experiments:

Hardware:

- 1 – Canon ZR60 Digital video camcorder
 - Canon AC adaptor
- 1 – Mini DV cassette
- 1 – VHS video cassette
- 1 – Portable desk/trolley
- 1 – Polaris Tool Interface Unit
- 1 – Polaris Position Sensor cable
- 1 – Polaris power cable
- 1 – Polhemus Fastrak Interface Unit
- 1 – Polhemus Fastrak Transmitter
- 1 – Polhemus Fastrak power supply/cable
- 1 – ATI force/torque sensor MUX box
- 2 – Serial port cables
- 1 – Logitech web cam
- 1 – 6' USB extension cord
- 1 – strain gauge power supply
- 1 – strain gauge instrumentation amplifier
- 2 – Tripods
- 1 – PC 2.4 GHz AMD Duron (tower, keyboard, mouse, monitor)
- 1 – Laptop PC (minimum 800MHz)
- 1 – Digital camera
- 1 – Mepore™*
- 1 – OpSite™*
- 1 – 2mm Allen key *
- 1 – 3mm Allen key*
- 1 – 4mm Allen key*
- 1 – small scissors*
- 4 – spare reflective balls for Polaris MDMA*
- 1 – Modified laparoscopic surgical tool (Maryland dissector)*
 - 3 – pieces of mounting bracket
 - 1 – Polhemus Fastrak Receiver
 - 1 – ATI Mini40 Force/Torque sensor
 - 1 – Polaris Position Sensor (MDMA)
 - Nuts and bolts for attachment and mounting

* Equipment requiring sterilization

Software:

Windows 2000
 Matlab 6.0 R12
 Tera Term Pro V2.3
 Logitech QuickCam V5.4.1
 FTGUI
 Matlab programs:
 - PMCS.m

Appendix A.3 OR Procedure

The following are the procedures for data collection in the OR for use of the experimental hybrid tool.

A.3.1 Pre-operative Set-up

Required Time: approximately 30min-60min with two people

Suggested start time: evening before surgery, or 0700h for 0800h surgical start

Set-up equipment as shown in Figure A.1.

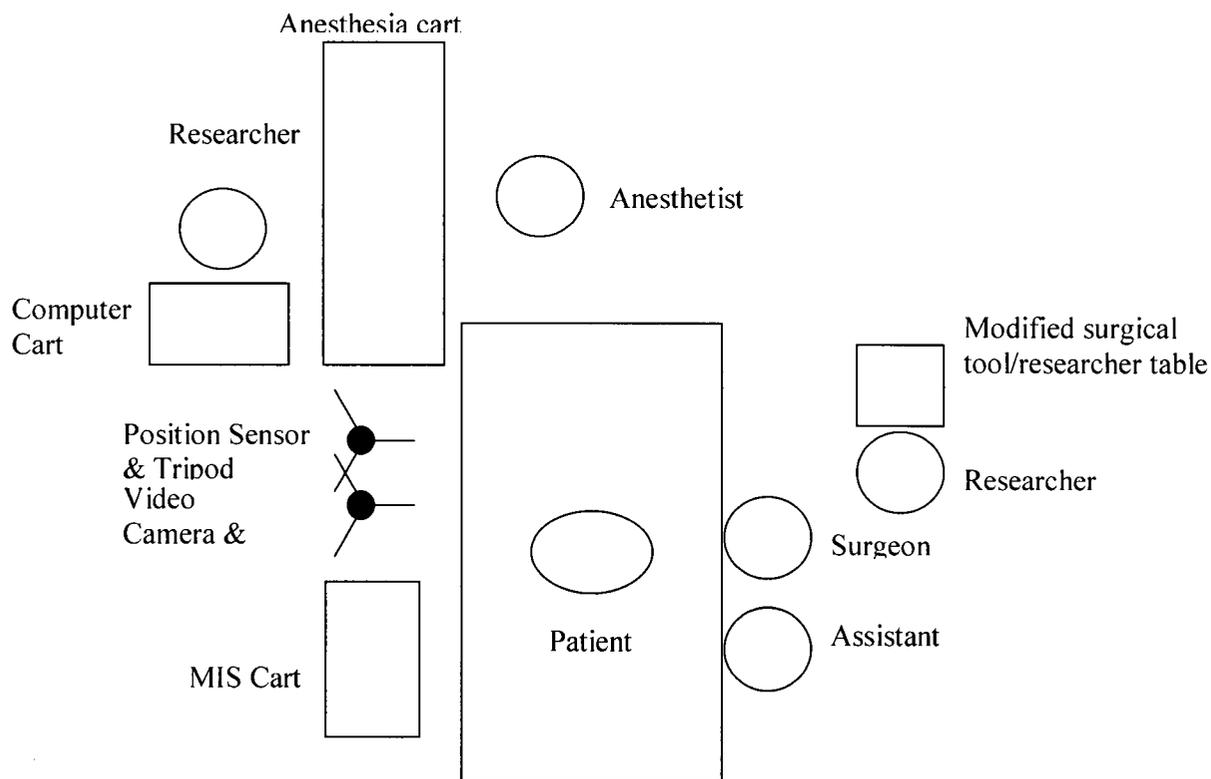


Figure A.1: University of British Columbia operating room experimental set-up

A.3.2 Pre-operative set-up and software initialization

Required Time: 10 - 30 min

Suggested start time: 7:00 for 8:00 procedure

A.3.2.1 Start-up Procedure Teraterm

- 1) Run Teraterm
- 2) Turn on Polaris Tool Interface Unit (wait 20 sec for beep – *RESETBE6F* will appear in the command window)
- 3) Type *COMM50000* – Reply: *OKAYA896*
- 4) Teraterm window: Setup – Serial Port : change baud rate to 115200
- 5) Type: *INIT_enter* (note: *_* means space bar) – *OKAYA896*
- 6) Teraterm Window: File – exit

A.3.2.2 OR Data Collection Procedures

- 1) Start digital video camcorder (focus camera on surgeon arm and experimental tool)
- 2) Start the MIS VCR to record the laparoscopy
- 3) Start Matlab R12
- 4) Set current directory to file where data is to be collected (i.e., *OR_test3*)
- 5) In command window type: *PMCS*
- 6) Graphical user interface will appear
- 7) Select radio buttons for all faces
- 8) STOP check equipment connections before initialization of Polaris
- 9) Press the *Initialize Polaris* button when ready – Polaris will beep in acknowledgment
- 10) Wait until all status bars are illuminated in yellow (Check for error messages in the Command Window)
- 11) When the surgeon indicates that she is ready to use the experimental tool the sensor cords are plugged in to the tool interface units under the surgical table.
- 12) On the laptop start *FTGUI*
- 13) In *FTGUI*, select *logging to port* radio button, and select appropriate folder
- 14) In *FTGUI*, select *continuous* data collection radio button
- 15) In *FTGUI*, push the *Options* button – output data – *Metric*
- 16) In *FTGUI*, push the *Options* button – *hemispheres* – set hemispheres to 0,0, -1
- 17) In *FTGUI*, push *Record Data*
- 18) In *PMCS*, select the appropriate tool used by the surgeon (Tool 2 for our case)
- 19) Monitor the status bars and adjust equipment if necessary (i.e. camera) if required. (Red status bar: missing marker, Yellow status bar: loading tool identification files, Green status bar: tracked marker)
- 20) At the end of the procedure in *FTGUI* press Stop data recording, press Stop Tracker – followed by Shut-down Polaris (MUST BE IN THIS ORDER)

Appendix B

Operational Definitions

B.1 Hierarchical Decomposition Operational Definitions (McBeth 2001)

A hierarchical decomposition modified from Cao by McBeth in 2001 was used to organize the OR data and also to find tasks that are analogous in the simulators and OR. The five-level decomposition describes the procedure in terms of surgical phases and stages, tool tasks and subtasks, and fundamental tool actions (Figure B.1). We are looking primarily at the Task and Subtask levels of the hybrid experimental surgical tool during dissection tasks.

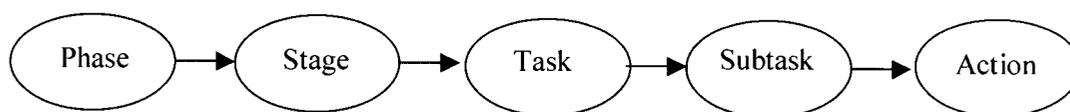


Figure B.1: *Five levels of the hierarchical decomposition*

B.1.1 Phase Level

Phases are the fundamental levels of a procedure forming the backbone and the foundation for further decomposition. A laparoscopic cholecystectomy procedure is divided into five distinct phases as shown in Figure B.2. Each phase has a particular goal associated with it to be accomplished in order to proceed to the next phase. This study dealt with Stages, Tasks and Subtasks in the cystic duct and gallbladder dissections only. Future work may be able to incorporate all aspects of the procedure by having multiple tool tips.



Figure B.2: *Five phases of a laparoscopic cholecystectomy*

B.1.2 Stage Level

The phase levels are further divided into stages, which have goals, but the goals do not have to be successfully completed before proceeding to the next stage. The stages of the cystic duct and gallbladder dissection (CDD and GBD) phases are shown below (Figure B.3) All stage level definitions are based on video observations.

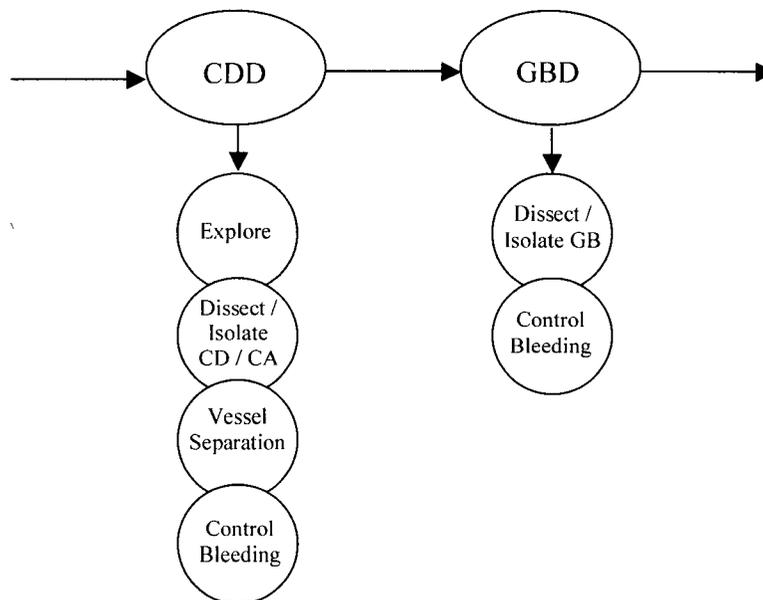


Figure B.3 – Stage level diagram for cystic duct dissection (CDD) and gallbladder dissection (GBD)

B.1.3 Task Level

A task is a set of movements performed with a single tool to achieve a desired effect. A number of tasks may be required to successfully complete a stage within a procedure. A task segment is defined from the time the tool tip is placed in the distal end of the trocar until the tool is pulled out through the same trocar. We chose the dissection task to investigate.

B.1.4 Subtask Level

The subtask level defines how the experimental tool tip is moving inside the patient. The subtasks are shown in Table B.1 for a dissection task.

Table B.1: *Hierarchical subtask dissection definition*

Subtask Name	Definition	Start	Stop
Free Space Movement: Approach*	Tool is moving toward tissue upon entry into the trocar	Entry of the tool tip into the distal end of the trocar	Tool tip in contact with tissue
Tissue Manipulation	Tool is in contact with the tissue	Initial contact of the tool tip with the tissue	Final contact of the tool tip with the tissue
Free Space Movement – Withdrawal	Tool is moving away from the tissue being pulled out of the trocar	Final contact of the tool tip with tissue	Exit of the tool tip from the distal end of the trocar

*tool moving in freespace (no tissue manipulation)

B.1.5 Action Level

The action states are made up of 12 types of distinct tool movements. The action level was not examined in this study. It is possible for tool movements to be a combination or a collection of actions. There are a total of 72 feasible combinations of the 12 action states (McBeth 2001).

B.2 Performance Measure Definitions

The definitions of the performance measures presented in Chapter 4 are shown in Table B.2. The individual component of the performance measures are calculated by projecting the tool path vectors onto the tool axis vectors, and allows us to compare performance measures across different settings.

Table B.2: Kinematics and force performance measures

Kinematics Measure	Performance Measure	Definition
Distance from mean	Absolute (mm)	$\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (z_i - \bar{z})^2}$
	Roll (radians)	$Roll_i - \overline{Roll}$
Velocity	Axial (mm/s ²)	\dot{z}_i
	Grasp (mm/s ²)	\dot{y}_i
	Translate (mm/s ²)	\dot{x}_i
	Transverse (mm/s ²)	$\sqrt{\dot{x}_i^2 + \dot{y}_i^2}$
	Absolute (mm/s ²)	$\sqrt{\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2}$
Acceleration	Roll (rad/s ²)	$Roll_i$
	Axial (mm/s ³)	\ddot{z}_i
	Grasp (mm/s ³)	\ddot{y}_i
	Translate (mm/s ³)	\ddot{x}_i
	Transverse (mm/s ³)	$\sqrt{\ddot{x}_i^2 + \ddot{y}_i^2}$
Jerk	Absolute (mm/s ³)	$\sqrt{\ddot{x}_i^2 + \ddot{y}_i^2 + \ddot{z}_i^2}$
	Roll (rad/s ³)	\ddot{Roll}_i
	Axial (mm/s ⁴)	$\ddot{\ddot{z}}_i$
	Grasp (mm/s ⁴)	$\ddot{\ddot{y}}_i$
	Translate (mm/s ⁴)	$\ddot{\ddot{x}}_i$
Force	Transverse (mm/s ⁴)	$\sqrt{\ddot{\ddot{x}}_i^2 + \ddot{\ddot{y}}_i^2}$
	Absolute (mm/s ⁴)	$\sqrt{\ddot{\ddot{x}}_i^2 + \ddot{\ddot{y}}_i^2 + \ddot{\ddot{z}}_i^2}$
	Roll (rad/s ⁴)	$\ddot{\ddot{Roll}}_i$
	Axial (N)	z_f
	Grasp (N)	y_f
Force	Translate (N)	x_f
	Transverse (N)	$\sqrt{x_f^2 + y_f^2}$
	Absolute (N)	$\sqrt{x_f^2 + y_f^2 + z_f^2}$
	Roll Torque (N-m)	y_i

Appendix C

University of British Columbia CREB approval

Appendix D

Medicine Meets Virtual Reality Conference Submission

This document was submitted to and presented in poster form for The 11th Annual Medicine Meets Virtual Reality Conference in Newport Beach, California, USA in January 2003.

Quantitative measures of transfer of training and validation of laparoscopic surgical simulators

Catherine Kinnaird¹, Joanne Lim¹, Antony J. Hodgson¹ PhD, Alex G. Nagy² MD, Karim Qayumi² MD PhD, Lance Rucker³ DDS, Karon MacLean⁴ PhD

Departments of Mechanical Engineering¹, Surgery²,
Oral Health Sciences³, and Computer Science⁴
University of British Columbia, Vancouver, BC, V6T 1Z4, CANADA

Abstract

Objective measures of surgical performance in minimally invasive surgery are of interest for students, surgeons and the public alike. Current assessments of surgical performance in the operating room are subjective and potentially unreliable (Rosser, 1998). Surgical simulators have been recognized as a potential source of objective assessment. However, until these simulators have been shown to be a valid and reliable measurement source, their use in surgical education remains minimal. The goal of this project is to use a multi-faceted approach to surgical assessment in the operating room, and to compare these measures to performance in analogous tasks on surgical simulators, both bench-top and virtual reality. In order to organize this research, a hierarchical decomposition of a laparoscopic cholecystectomy is used to divide the surgery into many component tasks (McBeth, 2002). The operating room assessment will utilize performance measures previously shown to be reliable such as time and postural data (McBeth, 2002), as well as incorporating a new measure of force/torque, and a new method to measure kinematics. The force/torque measures will use a similar approach as Rosen (Rosen, 2001), whereby a 3-dimensional load cell measures forces and torques on a laparoscopic tool. Postural data will be gathered using an optical tracking system. Finally, kinematics data will be gathered using a fusion of two types of sensors, optical and magnetic tracking systems. The optically tracked points will act as fixes, whereby the magnetic sensor data with its faster update rates will be fit to these optically tracked points.

One specific aspect of this project involves assessing the transfer of training from the simulator to the operating room. A control group and an intervention group, comprised of surgeons at various skill levels will perform surgical tasks in the operating room. The intervention group will then receive simulator training. Both groups will be re-tested in the operating room. Comparisons between the groups using the aforementioned performance measures will then be used to assess the transfer of training effects.

Validation of laparoscopic surgical simulators is yet another component of this project. We want to quantify the subjective measure of face validity using the performance measures described. Through research with expert surgeons in the operating room and the two types of simulators, bench-top and virtual reality, we plan to quantitatively assess simulator validity, as well as establish a method to effectively assess other surgical simulators.

The novelty of this research lies in the multi-pronged approach to quantitatively assess surgical performance in the operating room in order to validate surgical simulators. Many of these measures have been used alone in previous work, but this would be the first time they have been combined in this way, as far as we know. This work is done in conjunction with the Center of Excellence for Surgical Education and Innovation (CESEI), which is organized by the University of British Columbia and the Vancouver Hospital and Health Sciences Center. The mission of the CESEI is to provide multi-disciplinary academic educational center through the use of modern electronic technology. Validated simulators provide a potential source of training and certifying surgeons, as well as designing and evaluating tool designs.

References

McBeth, P. (2002). Thesis: A methodology for quantitative performance evaluation in minimally invasive surgery. University of British Columbia, Vancouver, BC, Canada.

Rosen J, Hannaford B, Richards CG, Sinanan MN. (2001). Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skill. *IEEE Transactions on Biomedical Engineering*. 48(5):579-91.

Rosser JC, Rsser LE, Savalgi RS. (1998). Objective evaluation of a laparoscopic surgical skill program for residents and senior surgeons. *Archives of Surgery*. 133(2):657-661.

Appendix E

Society of Gastrointestinal Endoscopic Surgeons (SAGES) Conference Submission

This document was submitted to and presented in poster form (by Dr. Hamish Hwang) for The 10th World Congress of Endoscopic Surgery in Denver, Colorado, USA in April 2004.

Objective Multi-Modal Surgical Performance Analysis

Joanne Lim¹, Catherine Kinnaird¹, Antony J. Hodgson¹ PhD,
Alex G. Nagy² MD, Karim Qayumi² MD PhD

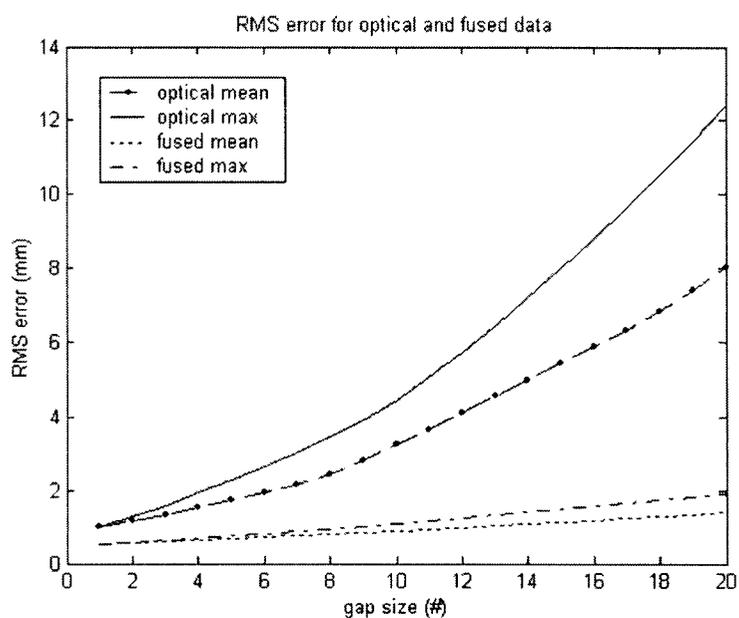
Departments of Mechanical Engineering¹, and Surgery²,
University of British Columbia, Vancouver, BC, V6T 1Z4, CANADA

Abstract

Objective measures of surgical performance in minimally invasive surgery are of interest for students, surgeons and the public. The goal of this project is to use a multi-faceted approach to surgical assessment in the operating room, and to compare these measures to performance in analogous tasks on surgical simulators.

The operating room assessment will use performance measures such as time, tool tip forces and torques (newly added), and tool kinematics. A commercial 3-D load cell mounted on a laparoscopic tool measures the forces and torques. Strain gauges are mounted onto the tool handle to measure surgeon grip levels. Kinematics data is gathered using both optical and magnetic sensors and the resulting data streams are fused to improve accuracy and reliability.

This data fusion will be done using a simple yet effective algorithm we have recently developed. The optical sensor data is regarded as extremely accurate, but it is subject to occlusion and has a comparatively low sampling rate. The magnetic data is acquired more frequently and is never occluded, so we fuse the magnetic data to the optical data for the entire data stream. This gives a complete set of data even when there are optical data gaps. As shown in the figure, the fused estimate is roughly 6-8X more accurate when optical data is missing than an estimate based on interpolating across the gap with optical data alone.



The novelty and uniqueness of this research lies in the multi-pronged approach to quantitatively assessing surgical performance in the operating room. Although some of these measures have been used individually in previous work, to our knowledge they have not previously been combined in this fashion, nor have tool tip forces been measured throughout a live surgery.

Appendix F

Transfer: of Training from Simulator to Operating Room

The original goal of this project was to study the issue of transfer of training from simulator to human operating room, as this is a subject that needs analysis in the surgical education and simulator arenas. But due to many logistical nightmares such as patient recruitment, scheduling, and many others, this project had to be converted to the study described in the manuscript.

F.1 Transfer of Training

The subject of transfer of training is a widely studied topic in many fields, not only in surgery. Likely the most widely known research in this area would be with flight simulators. The first and simple flight simulator was created around 1910 in France when a young student pilot could practice simple controls in a smaller modified type plane (Moore 2002). After many technological advances since that time, computer-based flight simulators have been commonly used in the training of pilots (Wentink 2003, Zeyada 2000). It is time the medical community also takes a closer look at using surgical simulators in the training and credentialing of surgeons.

Studies have been conducted outside of flight training and surgical training venues to really study what transfer of training is, and what affects this transfer. There is also a difference between learning and training transfer that many do not realize. True transfer of training occurs when the behavior transfers between two distinct and novel situations, while learning occurs when the behavior transfers between two identical situations (Auffrey 2001). There are also the concepts of near and far transfer; near transfer which is between nearly identical situations, and far transfer is between novel contexts. "True" training transfer is thought to occur quite rarely, and the teaching is thought to be most useful when it is specific, and practiced in an environment similar to the intended situation (Auffrey 2001).

To encourage successful learning and training transfer, it is necessary to vary the conditions of practice: part versus whole task methods (Auffrey 2001). Part methods focus on breaking the whole task into significant pieces, which are then practiced individually and explained in terms of how they fit in the whole task. Whole methods focus on the repetition of the task in its

whole. Whole methods are acceptable for simple tasks, while part methods should be used for individual difficulties or for complex time-consuming tasks (Teague 1994).

The conditions of transfer of training include the generalization of knowledge and skills acquired in training, and the maintenance of that learning over time (Ford 1997). There are three key factors that can impact training outcomes and transfer: 1) training design 2) trainee characteristics 3) work environment factors (Ford 1997). These researchers also see the need for multiple performance measures (other than self-report) for developing a more complete understanding of training transfer. It is reasonable to expect that individual's personality might affect future performance, but also affect the individual's enthusiasm to learn, learning strategies used, rate of skill acquisition, and of course, transfer of training. There are also environmental factors such as support, work climate, and opportunities that are important factors impacting training transfer (Ford 1997, Foxon 1993).

F.2 Assessing Transfer of Training

The assessment of the transfer of training from one environment to another is not a concept that is unique to surgical education. It has long been used in the flight training industry, as most if not all commercial pilots are trained and assessed in flight simulators. There have been very early evaluations of flight simulators that demonstrate training and cost-effectiveness (Flexman 1972, Williams 1949). This industry also takes advantage of a concept known as the "transfer effectiveness ratio (TER)" (Blaiwes 1984). This TER is typically 0.75, which shows a reduction of three hours of in-flight training is accomplished by 4 hours of simulator training.

F.3 Research Questions

We have formulated some research questions that revolve around our main objective.

Answering these questions will give us a better view on what we are trying to investigate.

F.3.1 Do novices who practice in simulators get better in the OR?

"See one. Do one. Teach one." This is the traditional surgical education mantra that was not far off from the truth. A surgical student would spend time with an experienced surgeon in the operating room observing and noticing the particulars of surgery. The next step would be to try out this surgery for oneself. And of course, logically, would be to teach the next batch of up and coming novice surgeons. As absurd as all this sounds, it is the way many surgeons have learned their trade. This obviously is not an acceptable method of education, and things have started to change. And most recently, surgical simulators have come to the forefront as a possible good method for surgical education.

F.3.2 What do we want to know?

Do novice surgeons who practice in simulators show a significant improvement in the operating room? If a novice surgeon spends X amount of time practicing on a simulator, will there be a quantifiable improvement in the operating room performance, as opposed to a similar novice who does not have any simulator training. So this would be the ultimate question, do novices who practice in simulators quantitatively improve their surgical performance in the operating room?

F.5.3 Why do we need to know?

Why is it important that surgical educators find out if training in a simulator transfer to the operating room? It has been shown that intra-operative assessments are subjective (Lentz 2001). In the simulator, the assessments are all objective and quantifiable. If a novice surgeon could do the majority of practice in a simulator, it is possible that many thousands of dollars would be saved in operating room expenses as has been mentioned earlier in this manuscript. It would be optimal for a novice to stay in the surgical simulator until all skills have been learned and practiced to an expert level, and then the novice would then be moved into the operating room. All the psychomotor and many cognitive skills would already be honed, and there would be less time required in the operating room to practice trivial basic tasks. In the simulator, the skills would all be objectively assessed as opposed to the subjective operating room assessments usually administered by the attending expert surgeon.

F.5.4 What we know

What do we know now? It has been shown that experienced surgeons are better in simulators than novices (McNatt 2001). And novices who practice in simulators do show improvement (Risucci 2001).

Seymour and associates, in a breakthrough study in the human operating room, have published one of the more recent and respected studies supporting the theory of transfer of training between simulator and operating room (Seymour 2002). They were one of the first groups to conduct a true transfer of skill study from simulator to clinical operating room. Surgical residents (PGY1-4) were randomly assigned into two groups: one group to receive virtual reality (VR) training in addition to standard training (ST), and the other group to receive only ST. All subjects completed a series of tests in visuo-spatial and perceptual abilities prior to training. Psychomotor abilities and VR training were tested on the Minimally Invasive Surgical Trainer-Virtual Reality (MIST-VR). All operative procedures were videotaped. Their

measurements were based on explicitly defined observable operative errors (8 defined errors) and length of time. They found that the ST group made six times as many of the defined errors when compared to the VR group. The ST group also spent more time completing the task, but it was not statistically significant. This study was one of the first to demonstrate that training in a simulator does correlate with improved performance in the human operating room.

Grantcharov and associates published the most recent study in the study of transfer of training to the operating room in 2004 (Grantcharov 2004). They investigated whether laparoscopic skills acquired in a VR simulator could be transferred to operations. This would therefore also validate the role of VR simulation as a tool for surgical skills training. The study participants consisted of 20 surgeons with limited laparoscopic cholecystectomy experience (from 0-8 median 4.5). All participants performed a baseline laparoscopic cholecystectomy under supervision of an experienced surgeon. The trainees were then randomized to receive either VR training or a control group with no additional training. The VR group trained on the MIST-VR doing 10 repetitions of all 6 tasks (of progressive complexity) available in the system. Within 14 days of the baseline laparoscopic cholecystectomy, all participants performed another laparoscopic cholecystectomy. These procedures were videotaped and assessed by two senior surgeons using predefined rating scales. The results show the VR trained group performed the laparoscopic cholecystectomy faster than the control group, and showed an improvement in error score and economy of movement. The limitations of this study as mentioned by the authors include the fact that the scoring of OR performance was subjective (although minimized by defining objective and easily assessed scoring criteria), and there was a small sample size. This study further supported the idea of transfer of training between simulator and operating room.