RELIABILITY STUDY OF THE LAPAROSCOPIC SKILLS INDEX (LSI): A NEW MEASURE OF GYNECOLOGIC LAPAROSCOPIC SURGICAL SKILL

by

Jerry Shime

A thesis submitted in conformity with the requirements for the Degree of Master of Science, Graduate Department of Health Administration, Division of Clinical Epidemiology and Health Care Research, Faculty of Medicine, University of Toronto

© Copyright by Jerry Shime 2000
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni desextraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.
RELIABILITY STUDY OF THE LAPAROSCOPIC SKILLS INDEX (LSI):
A NEW MEASURE OF GYNECOLOGIC LAPAROSCOPIC SURGICAL SKILLS

Jerry Shime

MSc in Clinical Epidemiology, 2000

Department of Health Administration, University of Toronto

ABSTRACT

This thesis deals with the construction and reliability testing of the LSI, a new multiitem measure of laparoscopic skills in gynecology to be used in the human model.

Methods

The construction of the LSI involved (1) item selection (2) selecting a method to scale responses (3) design (4) choosing a scoring method. Reliability was evaluated in two phases (1) a pilot study in which six raters reviewed six videotaped procedures (subjects), on two occasions, separated by one month (2) a main study in which four raters reviewed 20 videotaped procedures (subjects). The main study was conducted following efforts to improve the LSI reliability.

Results of the main study

Cronbach’s alpha was 0.95 indicating a high level of internal consistency. A “best measure” reliability coefficient for interrater reliability was ICC = 0.77 (95 percent CI 0.56 to 0.90) which is considered to be very good. This result was attained after excluding one rater where a systematic bias was evident.

The LSI appears to have the properties of a unidimensional index in which the item variables are true components of the overall attribute. Before entering into general usage, construct validity must be determined and the two methods of administering the LSI, in vivo and by videotape, must be compared.
ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. Claire Bombardier and the members of my thesis committee, Dr. Jean-Paul Szalai and Dr. Rose Kung for their assistance and advice throughout the development and completion of this project.

I would also like to thank my colleague and collaborator throughout this project, Dr. Richard Pittini in particular for his expertise in videocapture and videoediting.

Thank you to Dr. James Wright for providing me with one-on-one teaching during the directed reading course which was relevant to the completion of the thesis.

Thank you to Mr. Marko Katic for providing his technical expertise in generating the statistical output.

I wish to express my sincere appreciation to the following colleagues who assisted as content experts in the development of the LSI or as raters during the reliability studies: Drs Dalip Bhangu, Nancy Durand, Douglas Gare, Ellen Greenblatt, Rose Kung, Robert Hall (Orillia), Robert Hall (Toronto), Nicholas Leyland, Kay Lie, Titus Owolabi, Wusun Paek, Bruce Thomas, Fay Weisberg, and Arthur Zaltz.

I am grateful to the staff gynecologists at Women's College Hospital for their enduring support and for their participation in the various surgical procedures used in the study. Thank you as well to the residents who participated in the surgical procedures.

I would like to thank Dr. Veena Zaveri for her assistance in recruiting study subjects.

My special thanks to Ms. Joy Duran for her expert assistance throughout the project and, in particular, in the preparation of this manuscript.

Finally, I would like to thank my wife, Sheryl, for her unwavering support and patience throughout the years of study which culminated in the completion of this project.
# TABLE OF CONTENTS

**ABSTRACT** .................................................................................................................. i

**ACKNOWLEDGMENTS** ................................................................................................... iii

**TABLE OF CONTENTS** .................................................................................................... vii

**LIST OF TABLES** ........................................................................................................... viii

**LIST OF APPENDICES** ................................................................................................... ix

**CHAPTER 1: INTRODUCTION TO GYNECOLOGIC LAPAROSCOPY** ................................. 1

1.1 Definition of laparoscopy ................................................................................................. 1
1.2 Development of modern laparoscopy .............................................................................. 1
1.3 Technique of gynecologic laparoscopic surgery ............................................................ 2
1.4 The challenge of laparoscopic surgery ........................................................................... 3
1.5 Benefits of operative laparoscopy ................................................................................... 4
1.6 Disadvantages of operative laparoscopy ......................................................................... 4
1.7 Summary ......................................................................................................................... 7

**CHAPTER 2: TRAINING AND CREDENTIALING IN OPERATIVE LAPAROSCOPY** ............... 8

2.1 Guidelines for training and credentialing in operative laparoscopy ............................... 8
2.2 Level of competence in operative laparoscopy required for specialty certification in Obstetrics and Gynecology ......................................................................................... 9
2.3 Resident training in gynecologic operative laparoscopy ................................................. 9
2.4 Minimal access therapy training units ........................................................................... 10
2.5 Summary ......................................................................................................................... 12

**CHAPTER 3: MEASURING LAPAROSCOPIC SURGICAL SKILLS** ..................................... 13

3.1 Definitions ....................................................................................................................... 13
3.2 The assessment of clinical skills ..................................................................................... 13
3.2.1 The assessment of surgical technical skills ................................................................ 14
3.2.2 Objective structured assessment of technical skills (OSATS) .................................... 16
3.2.3 The objective assessment of laparoscopic technical skills .......................................... 18
3.2.4 The objective assessment of surgical skills in the human model ............................... 19
3.2.5 The objective assessment of laparoscopic skills in the human model ....................... 21
3.3 Summary ......................................................................................................................... 22

**CHAPTER 4: THE LAPAROSCOPIC SKILLS INDEX (LSI)** ............................................... 23

4.1 Introduction ...................................................................................................................... 23
4.2 Type of measurement ...................................................................................................... 23
4.3 Scale construction ........................................................................................................... 24
4.3.1 Item selection ............................................................................................................. 25
4.3.1.1 Item generation ................................................................................................... 26
### Chapter 8: Measures Taken to Improve the Design, Rater Preparation and Contents of the LSI

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1</td>
<td>Internal consistency</td>
<td>63</td>
</tr>
<tr>
<td>8.2</td>
<td>Reproducibility</td>
<td>63</td>
</tr>
<tr>
<td>8.2.1</td>
<td>Subject variability</td>
<td>64</td>
</tr>
<tr>
<td>8.2.2</td>
<td>Procedure variability</td>
<td>64</td>
</tr>
<tr>
<td>8.2.3</td>
<td>Rater variability</td>
<td>65</td>
</tr>
<tr>
<td>8.2.3.1</td>
<td>Process criteria</td>
<td>65</td>
</tr>
<tr>
<td>8.2.3.2</td>
<td>Conversion criteria</td>
<td>65</td>
</tr>
<tr>
<td>8.3</td>
<td>Summary</td>
<td>68</td>
</tr>
</tbody>
</table>

### Chapter 9: Main LSI Reliability Study ("Main Study")

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Main study research questions</td>
<td>69</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Primary research questions</td>
<td>69</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Secondary questions</td>
<td>69</td>
</tr>
<tr>
<td>9.2</td>
<td>Main study summary of methods</td>
<td>69</td>
</tr>
<tr>
<td>9.3</td>
<td>Main study subject (procedure) selection</td>
<td>70</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Main study procedure eligibility criteria</td>
<td>70</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Main study procedure exclusion criteria</td>
<td>70</td>
</tr>
<tr>
<td>9.4</td>
<td>Main study rater selection</td>
<td>70</td>
</tr>
<tr>
<td>9.5</td>
<td>Main study sample size calculation</td>
<td>71</td>
</tr>
<tr>
<td>9.6</td>
<td>Main study feasibility</td>
<td>72</td>
</tr>
<tr>
<td>9.7</td>
<td>Main study ethics approval</td>
<td>72</td>
</tr>
<tr>
<td>9.8</td>
<td>Main study video recordings and videotape compilations</td>
<td>73</td>
</tr>
<tr>
<td>9.9</td>
<td>Main study rating procedure</td>
<td>74</td>
</tr>
<tr>
<td>9.10</td>
<td>Main study data collection and entry</td>
<td>74</td>
</tr>
<tr>
<td>9.11</td>
<td>Main study outcomes</td>
<td>74</td>
</tr>
<tr>
<td>9.11.1</td>
<td>Internal consistency</td>
<td>74</td>
</tr>
<tr>
<td>9.11.2</td>
<td>Rater reliability</td>
<td>75</td>
</tr>
<tr>
<td>9.12</td>
<td>Main study results</td>
<td>75</td>
</tr>
<tr>
<td>9.12.1</td>
<td>Management of various data issues prior to statistical analysis</td>
<td>76</td>
</tr>
<tr>
<td>9.12.1.1</td>
<td>Observational conflicts</td>
<td>76</td>
</tr>
<tr>
<td>9.12.1.2</td>
<td>False positives</td>
<td>77</td>
</tr>
<tr>
<td>9.12.1.3</td>
<td>False negatives</td>
<td>77</td>
</tr>
</tbody>
</table>
9.12.2 Analysis options ........................................... 77
9.12.3 Internal consistency ........................................ 79
9.12.4 Rater reliability ........................................... 80
9.13 Summary ....................................................... 83

CHAPTER 10: DISCUSSION ........................................... 84
10.1 Introduction ..................................................... 84
10.2 Prevention and management of bias in the LSI reliability studies ........................................... 86
10.3 Limitations of the research ..................................... 88
10.4 Sensibility of the LSI ........................................... 91
10.5 Future research .................................................. 93
10.6 Conclusions ...................................................... 95

REFERENCES .............................................................. 96
LIST OF TABLES

Table 1: Background Information on Content Experts ........................................... 34

Table 2: Corrected Item-Total Correlations across All Raters for Time 1 and Time 2 .......................................................... 57

Table 3: Cronbach's Alphas for Individual Raters and Overall for Time 1 and Time 2 .......................................................... 57

Table 4: Corrected Item-Total Correlations for Individual Raters, Pilot Study ..... 58

Table 5: Intrarater Reliability (ICC) and 95 % confidence intervals for Individual Raters .......................................................... 59

Table 6: Interrater Reliability (ICC) for Individual Items and Overall Scores for Time 1, Time 2, and Times 1 and 2 Combined and 95 % Confidence Intervals 60

Table 7: Mean, Standard Deviation and Range of Total Scores for Individual Raters .......................................................... 78

Table 8: Mean and standard deviation scores for individual items ..................... 79

Table 9: Corrected Item-Total Correlations and Cronbach's Alphas for Individual Raters and for All Raters Combined (Main Study) .......................... 80

Table 10: Interrater Reliability (ICC) and 95% Confidence Intervals for Different Combinations of Raters and Subjects (Main Study) ..................... 81

Table 11: Individual Item ICC's and 95% Confidence Intervals for Different Combinations of Raters and Subjects (Main Study) ..................... 82
LIST OF APPENDICES

Appendix 1  Guidelines for Resident Training in Endoscopic Procedures SOGC .... 104
Appendix 2  University of Toronto Department of Ob/Gyn ITER: page 10 of 11 pages ................................................................. 106
Appendix 3  Information Sheet and Consent Form: Pilot Study .................... 107
Appendix 4  Rater Instruction Sheet: Pilot Study ......................................... 109
Appendix 5  Item Descriptors Rater Information Package: Pilot Study ........... 110
Appendix 6  LSI Scoring Form: Pilot Study .............................................. 111
Appendix 7  LSI Rater Instruction Form: Main Study ................................. 112
Appendix 8  LSI Item Descriptors: Main Study ........................................ 113
Appendix 9  LSI Index Scoring Form: Main Study ..................................... 115
Appendix 10  Information Sheets and Consent Forms for Surgeons and Patients: Main Study .................................................. 117
Appendix 11  Ethics Board Approval ......................................................... 122
Appendix 12  Memorandum to staff re planned LSI Main Reliability Study .... 123
Appendix 13  Individual Rater Trend Lines for Total Scores (percent) Assigned to Each Subject: Main Reliability Study ..................... 124
CHAPTER 1: INTRODUCTION TO GYNECOLOGIC LAPAROSCOPY

1.1 Definition of laparoscopy

*Laparoscopy* is a procedure in which the abdominal cavity is examined with an endoscope inserted through any point in the abdominal wall.

1.2 Development of modern laparoscopy

Modern laparoscopy was made possible in the 1950's with the invention by Professor H. Hopkins at the University of Reading of a rod lens system which, coupled with fiberoptic cables could transport light into body cavities from an external source.\(^1\) Initially laparoscopy was used only for diagnosis and consisted of visualizing and occasionally biopsying intraperitoneal structures. In the 1960's, Raoul Palmer of Paris and Hans Frangenheim of Konstanz extended its use to include simple surgical techniques, namely, female sterilization and aspiration of ovarian cysts. Patrick Steptoe, who pioneered in vitro fertilization, published a book on gynecologic laparoscopy in 1967 and this resulted in the wide dissemination of knowledge about laparoscopy into the English speaking countries.\(^2\) In 1979, in France, Bruhat was the first to perform a laparoscopic operation using laser dissection but this modality was initially neglected in favour of electrosurgical techniques.\(^3\) Laser laparoscopy reached North America in the early 1980's and by the late 1980's the types and numbers of gynecologic procedures performed by laparoscopic surgery escalated rapidly. It is estimated that 80 percent of gynecologic operations that previously required *laparotomy*, defined as entry into the peritoneal cavity through an open abdominal incision, can now be performed by laparoscopic surgery.\(^1\)

Rapid improvement in the technology, including high powered light sources and triple chip cameras attached to the laparoscope, now make it possible for surgeons to perform laparoscopic procedures by viewing magnified, clear images of the abdominal cavity transmitted onto
strategically placed video monitors.

1.3 Technique of gynecologic laparoscopic surgery

Through a small incision in the margin of the umbilicus a 1.5 mm spring-loaded needle, known as the Veress needle, which connects to a carbon dioxide insufflator, is inserted into the peritoneal cavity. The peritoneal cavity is then distended with carbon dioxide creating a pneumoperitoneum. The Veress needle is removed and the laparoscope is then inserted through the same umbilical incision. Other small incisions, usually three and occasionally four, and ranging from 5 to 12 mm in length, are created in the left and right lower quadrants of the abdomen for insertion of various trocar/cannulas. The trocars are removed to enable a variety of laparoscopic surgical instruments to be passed through the cannulas into the peritoneal cavity. These instruments include atraumatic and toothed grasping forceps, scissors, needle holders, suction/irrigation probes, coagulating and electrosurgical instruments, endoscopic stapling devices and specimen retrieval bags. Laparoscopic instruments are usually 30 cm long and are manipulated by the surgeon’s hands which are outside the abdominal cavity. The carbon dioxide laser and camera are connected directly to the laparoscope. The surgeon’s field of vision is controlled by movement of the laparoscope and by the appropriate use of zoom and panorama techniques. The images captured by the camera are displayed on two-dimensional video monitors. The surgeons perform procedures by viewing the monitor images. Dissection may be performed using the laser, electrosurgical tools, or by traditional methods using instruments modified for laparoscopy. A suction/irrigator is essential for removing plume, clearing debris, exposing bleeder and, frequently, for dissection. Suturing and knot tying are not employed in the majority of gynecologic laparoscopy operations. However, if necessary, knot tying can be performed in a variety of ways including: intracorporeal ligation, extracorporeal ligation with delivery of the knot to the desired site by specially designed instruments, and by
application of manufacturer preformed ligatures (Endoloops). Excised tissue can be removed in various ways depending on the size and consistency of the specimen. Specimen removal through the small trocar sites often entails tissue morcellation. Spillage of contents from a cystic structure can be avoided by inserting specially designed retrieval bags into the peritoneal cavity into which the specimens are placed, decompressed and then removed.

1.4 The challenge of laparoscopic surgery

Competence in the performance of laparoscopic surgery requires the acquisition of a variety of skills which differ from traditional surgery. Surgeons must learn to operate with long instruments which amplify tremor and are harder to control than conventional instruments. The instruments are limited in their range of motion by the small incisions through which they pass and their successful use requires mastery of the fulcrum effect created by the abdominal wall. The constraints imposed by the length and width of these instruments have limited the scope of engineering design. Optical differences between laparotomic and laparoscopic surgery also exist. Surgeons must learn to operate while looking at a monitor which is remote from the operative site. Conventional laparoscopic imaging systems supply two dimensional vision. Thus, depth perception is lacking and surgeons look for other cues to provide a sense of depth including the sense of touch and the interaction of light and shading. There are higher demands for ambidexterity as seen, for example, in the two-handed coordination required to integrate the actions of the suction/irrigator and the laser/laparoscope. New ligation skills, such as extracorporeal tying, must be learned. Unlike traditional laparotomies, where tissue is removed manually by the surgeon whose hands are physically within the operative field, specimen removal by morcellation and retrieval bag techniques is often time-consuming and challenging and may fatigue the less skillful operator.
1.5 **Benefits of operative laparoscopy**

Operative laparoscopy is frequently called “minimally invasive surgery” or “minimal access surgery” because of the small incisions used. Compared with traditional laparotomies, the benefits of operative laparoscopy include smaller incisions, less blood loss, less postoperative morbidity and drug usage, less scar formation, shorter length of hospital stay, earlier return to work and reduced need for sickness benefits\(^5,6,7,8,9,10\). During laparoscopy, the magnification of structures on the video monitor permits more accurate dissection, ready identification of bleeding vessels, more effective hemostasis and better access to some areas, such as the retrotuetine cul-de-sac where spatial limitations favour the use of long, slender laparoscopic tools or laser beam dissection over the traditional laparotomic approach\(^12,13,1\).

1.6 **Disadvantages of operative laparoscopy**

There are three primary disadvantages to a laparoscopic approach:

1. Procedures often take longer and therefore reduce the number of cases that can be scheduled on a particular day. This problem diminishes as the operator gains more experience. Also, the longer duration of surgery is offset by the advantage of shorter hospital stay.

2. Some procedures cannot be accomplished laparoscopically. For example, in gynecology disseminated ovarian cancer cannot be properly debulked using a laparoscopic approach. Also, very large, solid specimens, such as massive uterine fibroids, are often better removed by laparotomy, although newer electrical morcellation instruments are making a laparoscopic approach more feasible.

3. Complications of operative laparoscopy - the list of publications pertaining to surgical complications of laparoscopy is extensive\(^14,15,16,17,18,19\). In 1998, the Medical Protection Society of the United Kingdom reviewed 337 medicolegal claims across a range of surgical...
specialties. Fifty-three percent of the surgical claims related directly to the operative procedure and the specific area of concern highlighted was "problems with technique in minimally invasive surgery." 20

Complications specifically related to the technique of laparoscopy include injury secondary to Veress needle and trocar insertions, laser or electrosurgical thermal damage and mechanical injury from improperly applied endoscopic staplers.

The major sites of injury are:

1. Abdominal wall - Injuries include subcutaneous emphysema from misdirected carbon dioxide penetration of the epigastric vessels, and incisional hernias at 10 or 12 mm trocar sites when the underlying fascia is not closed14.

2. Retroperitoneal blood vessels - Injury to a major retroperitoneal vessel (terminal aorta, inferior vena cava or iliac vessels) from the Verres needle or trocars is reported to occur in 3-10/10,000 laparoscopies1.

3. Gastrointestinal tract - The gastrointestinal tract may be injured by the Veress needle or trocars, mechanical instrumentation or thermal energy. The incidence of bowel injury is estimated to be in the range of 0.5-1.8/1,000 laparoscopies16.

4. Urinary tract - The usual reported incidence of urinary tract injury is 0.2/1,000 laparoscopies. However, a recent series in Finland reported on 62,379 hysterectomies in that country over a five year period15. During laparoscopic hysterectomy ureteral injuries occurred in 13.9/1,000 cases compared with 0.4/1,000 for laparotomic hysterectomy and 0.3/1,000 for vaginal hysterectomy. Similarly, bladder injuries occurred in 2.2/1,000 laparoscopic hysterectomies compared with 0.1/1,000 and 0.2/1,000 laparotomic and vaginal hysterectomies respectively.
Nduka (1994) reported an overall incidence of electrosurgical injury during laparoscopy of 1/1,000-2/1,000\(^{21}\). The majority of injuries were unrecognized at the time of surgery and presented three to seven days postoperatively. The main causes of electrosurgical injuries were: inadvertent touching or grasping of tissue during current application, direct coupling between a portion of the intestine and a metal probe touching the activated probe, insulation breaks in the electrodes, direct sparking to the intestine from the diathermy probe and current passage to the intestine from recently coagulated, electrically isolated, tissue. The majority of injuries were caused by monopolar, as opposed to bipolar, diathermy.

In a 1998 review of 29,966 laparoscopic gynecology cases performed in seven French centres over nine years the mortality rate was 3.33/100,000 laparoscopies\(^{22}\). The overall complication rate was 4.64/1,000 and the rate of complications requiring conversion to laparotomy was 3.2/1,000. Thirty-four percent of complications occurred during access procedures (Veress needle and trocar insertions) and 28.6 percent of complications were not recognized during the operation.

Attempts have been made by different surgical specialties to quantify learning curves for laparoscopic procedures\(^{23,24,25,26,27,28}\). Agachan (1996) assessed various intraoperative and postoperative complications associated with laparoscopic colorectal surgery, chronologically classified into three groups. The total complication rate decreased from 29 percent in the first period to 11 percent by the second period (\(p<0.04\)) and 7 percent during the third period (\(p<0.05\)). The learning curve appeared to have exceeded 50 cases. In another study 86 surgeons participated in three day courses in laparoscopic cholecystectomy\(^{26}\). Each step of the operation was recorded by the instructor for successful performance, operative time, method of dissection and laser or cautery use. By the end of three days gall bladder dissection without perforation had not been mastered and
the authors concluded that the flat portion of the laparoscopic cholecystectomy training curve had not been reached at the program's end. A prospective survey of 181 urologic surgeons evaluated the relationship between laparoscopic complication rates and surgeon-dependent variables following a laparoscopic training course. At three months, surgeons who performed clinical procedures without additional training were 3.39 times more likely to have at least one complication compared with surgeons who sought additional training (p=.03). At 12 months, surgeons who had attended the training course alone or were in solo practice or performed laparoscopic surgery with a variable assistant were 4.85, 7.75, and 4.8 times more likely respectively to have had a complication than their counterparts who attended the course with a partner, were in group practice, or operated with the same assistant. (p=.004, p=.0008 and p=.0015, respectively). The authors concluded that the rate of complications during the clinical learning curve would be decreased by additional training following the initial course in laparoscopy and by an ongoing clinical association with surgeons performing similar procedures.

1.7 Summary

Technological advances have brought operative laparoscopy into the main stream of gynecologic surgery. The advantages of this “minimal access” surgery are numerous but new and challenging skills are required which lead to a prolonged learning curve during which the incidence of technique-specific complications is high.
CHAPTER 2: TRAINING AND CREDENTIALING IN OPERATIVE LAPAROSCOPY

2.1 Guidelines for training and credentialing in operative laparoscopy

The rapid dissemination of this new technology into practice has been catalyzed by media coverage glamorizing "laser surgery", intensive marketing by instrument manufacturers, and a push by government and hospital administrators to reduce hospital lengths of stay. With increasing frequency patients are demanding a laparoscopic approach and surgeons place pressure on hospitals for privileges to perform operative laparoscopy. In the first half of the 1990's it was common for certified gynecologists to attend brief (two to three day) laparoscopy courses consisting of didactic sessions, followed by "hands-on" experience with synthetic models and tissue specimens and a half day of training in the animal laboratory, usually partnered with another inexperienced laparoscopist. These gynecologists would then return to their own centres to perform operative laparoscopy with no further training. During the "learning curve" operations would take longer, complications were more frequent and many frustrated surgeons would abandon laparoscopic surgery in favour of traditional laparotomic approaches. More recently various organizations have attempted to establish guidelines outlining the need for surgeons to have didactic and laboratory experience, followed by a preceptorship prior to performing procedures independently. These guidelines share the following recommendations:

1. Certification in a surgical specialty.

2. Experience with basic laparoscopic techniques, such as diagnostic laparoscopy and laparoscopic sterilization

3. Completion of at least one accredited operative laparoscopy course consisting of i) didactic sessions which include instruction in basic skills, procedures, laser physics and electrosurgical principles ii) hands-on training using inanimate, tissue and animate models
A preceptorship during which the surgeon operates as the first assistant to an experienced preceptor after which the preceptor acts as the first assistant in a series of increasingly complex procedures. Subsequently, the preceptor may recommend to the department head that procedure privileges be granted.

2.2 Level of competence in operative laparoscopy required for specialty certification in Obstetrics and Gynecology

In 1994, the Royal College of Obstetricians and Gynaecologists in the United Kingdom classified laparoscopic procedures into four levels of difficulty. All MRCOG trainees would be expected to be competent in levels I and II. Gynecologists wishing to perform level III procedures were expected to register with the RCOG and attend an RCOG recognized advanced laparoscopy course followed by a preceptorship. The preceptorship would include critical analyses of the trainee’s videotaped operative procedures. Level IV procedures were to be performed only by subspecialists, such as urogynecologists and gynecologic oncologists, or in tertiary endoscopic surgery referral centres.

In 1995 the Society of Obstetricians and Gynaecologists of Canada produced guidelines for resident training in endoscopic procedures. Gynecologic laparoscopy was classified into three levels of difficulty. Levels I and II were considered to be minimum requirements for all specialty certification (Appendix 1). Residents would be encouraged to participate in the more advanced subspecialty level III procedures but independent performance of these operations would generally require extended training in laparoscopic surgery, usually within one of the subspecialties.

2.3 Resident training in gynecologic operative laparoscopy

Traditionally, gynecologic surgery is taught in a five year post graduate programme with residents acting as apprentices. During their training, neophyte gynecologic surgeons first have the
opportunity to observe the more experienced surgeons. Having gained experience as observers and as assistants, trainees then perform various aspects of the procedures under direct supervision until they have acquired the skills necessary to operate independently in a safe, effective manner.

In the area of operative laparoscopy there have been barriers to this traditional mentoring system. First, many of the faculty themselves, particularly the more senior members, have not trained in and do not perform operative laparoscopy. In some centres this has limited the faculty teaching pool which would be required to establish a successful training program in operative laparoscopy. Second, faculty surgeons undergoing a preceptorship as a prerequisite to hospital credentialing may compete with residents to perform cases. Third, experienced laparoscopic surgeons often call upon one of their colleagues for assistance in a complex case rather than rely on a trainee with lesser experience. Fourth, during surgical training there is a correlation between the complexity of the case and the level of resident training, and this is particularly evident in operative laparoscopy where the experience has shifted away from junior level to senior residents and attending surgeons. Finally, while laparoscopic procedures often achieve cost savings through reduced length of hospital stay, the actual operating time, especially for more major procedures such as laparoscopic assisted vaginal hysterectomy and morcellation and removal of large uterine fibroids, is often considerably longer than the traditional approach. Thus, teaching laparoscopic skills on living patients can lengthen operating time to an unacceptable degree for surgeons who are faced with increasing constraints in surgical scheduling.

2.4 Minimal access therapy training units

In order to overcome the barriers to learning operative laparoscopy several postgraduate surgery programs have developed training models outside the human operating theatre. These surgical skill laboratories incorporate laparoscopic trainers, pelvic trainers designed for
gynecology, video monitors and interactive, computer assisted learning work stations. The "specimens" vary from simple synthetic models to sophisticated, anatomically accurate, latex models, animal tissues and intact animals. The curricula focus primarily on practice skills including dominant, nondominant and two handed manipulation, eye-hand coordination, laparoscopic dissection, hemostasis and suturing. The aim is to provide trainees with the knowledge and skills necessary for laparoscopic surgery before entering the operating room.

Recently, virtual reality simulations have become more available with the arrival of low cost, industrial standard, multimedia computers and high performance graphic hardware, coupled with the spread of VR modeling and run-time software, low cost and free resources from the Web. In the past two years a minimally invasive surgical trainer (MIST) system has been developed covering a wide range of techniques from cholecystectomy to thoracic surgery. The software constructs a virtual environment screen showing the position and movement in real time of two laparoscopic instruments electronically linked to a high powered computer. The program is designed to improve surgical skills, motivate trainers and provide a pressure-free training environment, while enabling the tutor to identify when a particular trainee is ready to assist in the operating room by using objective standards of accomplishment. Experience with the MIST system is still very limited. To date it has not been used for gynecologic laparoscopy.

The concept of learning hierarchies refers to the necessity of learning the basic component skills of a routine before progressing to the full routine. This approach emphasizes repetition which, according to cognitive theories, facilitates the subsequent performance of motor activities by permitting more efficient interpretation of proprioceptive, visual and tactile feedback. Neurophysiologic testing on surgical residents failed to demonstrate a correlation between surgical skill and pure motor abilities such as speed and precision. Instead, surgical skill ratings of trainees
by faculty correlated with complex visual organization, somatosensory memory, and stress tolerance. *Visuospatial perceptual skill* refers to the ability to use landmarks to create a mental picture of relationships in three dimensional space. *Somatosensory memory* refers to the ability to interpret the sensory cues based on prior experience. *Stress tolerance* refers to the ability to distinguish essential detail from nonessential detail. Minimally invasive therapy units incorporate these concepts into curricula that permit novice surgeons to enhance proprioceptive and tactile perceptions while developing visuospatial orientation through an operative video camera. Frequent and repetitive practice is important to the retention of these skills. The importance of somatosensory memory, rather than pure motor ability, underlies the necessity of incorporating fresh tissue models that resemble true surgical tissues. Stress tolerance is enhanced not only by improving perceptual skills, but also by providing a thorough knowledge of laparoscopic equipment as well as their capabilities and limitations.

### 2.5 Summary

Guidelines which define the competency requirements for specialty certification in Obstetrics and Gynecology have been established, as have postcertification preceptorship guidelines for staff gynecologists seeking privileges to perform operative laparoscopy. Various barriers have limited the teaching of operative laparoscopy within the human setting. As a result, minimal access therapy training units have been established which use bench simulations for learning the basic skill components prior to assisting or performing procedures on patients.
CHAPTER 3: MEASURING LAPAROSCOPIC SURGICAL SKILLS

3.1 Definitions

_Skill_ is defined by the Unabridged Random House Dictionary as “expertness, practiced ability, facility or dexterity to do something well”\(^4\). A person who demonstrates technical skill is skilled in or familiar in a _practical_ way with a particular art, trade, or science\(^5\).

Clinical skills refer to the ability to do an appropriate history and physical examination, problem solve, arrive at a working diagnosis, and outline a plan of management\(^6\).

_Competence_ is defined by the Unabridged Random House Dictionary as “the quality of being competent: possession of required skills, knowledge, qualification and capacity”. Surgical competence implies an individual who has all the abilities required to undertake the duties required of a consultant surgeon to an agreed, acceptable standard. These abilities include knowledge, attitude, communication and technical skills, which, if properly integrated, result in sound judgement in decision making\(^7\). This thesis deals specifically with the assessment of operative laparoscopy technical skills in gynecology which is only one attribute of the larger concept of gynecologic surgical competence.

3.2 The assessment of clinical skills

Traditionally clinical skills have been “measured” by examinations, ward assessment forms and other forms of written examination. These methods have been shown to be lacking in either reliability or validity\(^8\). Multiple choice examinations are reliable but test factual recall. They lack content and construct validity for assessing higher cognitive skills such as problem solving\(^9,10\). Oral examinations have face validity but varying levels of reliability ranging from 0.19 to 0.80\(^11,12\). Ward assessments are highly subjective, skewed towards superlative ratings and are unreliable, with median ratings of several studies ranging from 0.25 to 0.37\(^13,14\).
During the past two decades efforts have been made to develop methods of assessing clinical skills which are more valid, efficient, and reliable than traditional techniques. These methods include written and computer-based patient management problems (PMP's), standardized patient (SP)-based examinations, and multistation examinations like the objective structured clinical examination (OSCE)\(^\text{33}\). The result has been a steady shift from subjective measures to objective measures. Objectivity refers to the comprehensive goal of value-free measurement. Objectification refers to the strategies employed to achieve an apparently judgement-free instrument. An instrument which has been constructed according to these strategies is one which has been objectified\(^\text{34}\).

The OSCE, first developed to evaluate clinical competence in medical students\(^\text{55}\), consists of multiple stations at which each candidate, faced with a discrete clinical task, is evaluated in an objective and structured way. The task and the assessment are standardized to ensure reliability. The OSCE provides reliability and validity which is superior to other traditional methods of evaluating clinical performance\(^\text{56}\). Several large studies of OSCE's administered in different countries showed rater reliabilities ranging from 0.59 to 0.93 which exceeded the available reliability measures for oral examinations and clinical ward assessments\(^\text{57}\). In a longitudinal study, modifications over time improved the reliability of one OSCE from 0.73 and 0.5 to 0.81 and 0.65. for the overall tests and the components, respectively\(^\text{58}\).

### 3.2.1 The assessment of surgical technical skills

It has been stated that 75 percent of the skills required by a surgeon are intellectual and personal and 25 percent involves technical competence, which includes manual dexterity and a capacity for focused and sustained attention\(^\text{59}\). As arbitrary as this division is, both are critically important. It is no good being able to handle tissue with great delicacy or suture beautifully if you are not sure when and where to do it; equally, it is not much use knowing what needs to be done if
it cannot be done skillfully. Unfortunately the objective assessment of cognitive knowledge in surgical trainees correlates poorly with the assessment of surgical skills by experienced senior surgical faculty ($r=0.267$). This suggests that clinical performance and cognitive knowledge measure different characteristics essential for clinical competence. Without a satisfactory objective measure of surgical skill no statement can be made about the technical ability of graduating surgeons nor can appropriate measures for improvement be initiated and evaluated.

Different models of surgical skills assessment have been proposed. One of these, outcome audits, has no place in the assessment of surgical trainees since long periods of study are required to produce statistical reliability during which patients may be receiving less than the best treatment; moreover, audits reflect the performance of the entire surgical and perioperative team as opposed to the individual surgical trainee. Procedure lists (trainee log books), in which a surgical trainee records each procedure that he/she has participated in and the level of participation, describe the range and volume of the trainee's experience but they are poor substitutes for the assessment of technical competence. Most commonly, in the surgical specialties, the technical competence of a surgical trainee is registered by a single item, global assessment within a multidimensional Intraining Evaluation Report (ITER) (Appendix 2). The grading is usually assigned by the unit post graduate resident supervisor at the end of a rotation. There are several problems with the ITER (1) it is a subjective exit evaluation provided "after the fact" as a resident leaves a service (2) it depends on staff recollection of the trainee's performance (3) faculty surgeons are rarely trained as raters and there is very little interrater consistency (4) there is an imbalance in the nature of comments favouring the use of superlatives (5) a single comment global assessment lacks detail regarding specific areas of strength and/or weakness and is therefore unhelpful in terms of targeting areas for improvement.
In an earlier attempt at objectifying the measurement of operative skills, Kopta (1971) defined surgical dexterity as a psychomotor skill in which a mental picture or perception is initially formed and then integrated with motor activity until eventually, the skill becomes "locked in" or autonomous. He identified seven basic skills in orthopedic procedures and provided checklists to evaluate each item on a score from 0 to 2. Two orthopedic surgeons then observed six different surgical procedures performed by six surgeons, ranging from intern to board certified orthopedist. The correlation coefficient between the two raters was \( r = 0.96 \). This result was likely inflated because the surgeon raters were aware of the training level for each surgeon subject and reliability assessment was a measure of consistency in rank ordering as opposed to agreement (see correlation and agreement, 5.4.1).

Using the instrument designed by Kopta, Delaney et al. (1978) evaluated four surgical procedures performed by different surgical trainees using three raters. The interrater agreement was 0.82. The authors concluded that the high interrater agreement was due to the large number of items on the skills checklists and the increased familiarity with the instrument on the part of the raters.

3.2.2 Objective structured assessment of technical skills (OSATS)

The Surgical Education Group at the University of Toronto developed a bench model examination for surgical skills called the OSATS. This has been adapted and modified by the Department of Obstetrics and Gynecology at the University of Toronto. The examination uses bench model simulation of surgical tasks rather than entire procedures to assess performance. Residents rotate from task to task, performing technical skills while being observed by trained surgeons. The examination is modeled after the OSCE. Faulkner et al. (1996) examined the concurrent validity of the OSATS. Twelve surgical residents at the University of Toronto were
ranked within level of training according to their OSATS mark and by surgical faculty. Correlations were generally high for the senior residents (0.66 to 0.71) but low for the junior residents. However, even for senior residents there was a small negative correlation between faculty ranking and OSATS rankings for the bench-check list scoring system. The authors concluded that scores on the OSATS accurately reflect the independent opinions of faculty regarding the technical skills of senior residents, suggesting that it is a valid measure of technical skill for these individuals. It was unclear whether the poor correlation for junior residents was a failure of the OSATS or the faculty rankings.

Martin et al (1997) compared the reliability of three scoring systems (check lists, global rating forms and pass/fail judgements for both live and bench formats) and the construct validity of the OSATS. Twenty surgical residents at the University of Toronto were tested. The disattenuated correlations between live and bench scores were high (0.69 to 0.72). Mean interrater reliability across stations ranged from 0.64 to 0.72 and internal consistency was moderate to high (alpha 0.61 to 0.74) for the live format using the check list, and for the live and bench formats using global ratings. Global ratings discriminated between resident levels for both formats (bench. $F(2,17) = 4.45, p<0.05$; live, $F(2, 17) = 3.55, p<0.05$); check lists did not. The study suggested that OSATS could reliably and validly assess surgical skills. Global ratings appeared to be a better method of assessment than task-specific check lists. Bench model simulation gave equivalent results to the use of live animals for this test format.

Reznick et al (1997) reported on a large scale administration of the OSATS aimed at investigating reliability and validity. Interstation reliabilities in the 0.80 range were achieved for this examination. This level of reliability was considered satisfactory by the authors for using examination scores to make high level decisions.
3.2.3 The objective assessment of laparoscopic technical skills

To date, the objective evaluation of laparoscopic skills has been conducted on bench and animal models within the context of minimal access therapy training units. Training performance has generally been measured using scoring systems which reward precision and/or speed in performing series of structured tasks or drills.\textsuperscript{37,12,70}

Cundiff (1997) evaluated the effectiveness of the laparoscopic skills program at Duke University Medical Center by comparing the performance of tasks (skills) at the beginning and end of a seven week program.\textsuperscript{37} Six post graduate year (PGY)1’s, six PGY3’s, and six PGY4’s all showed improvement in skill performance time in all categories (p=0.04 to p<0.001). Interestingly, while the residents’ self assessment of competence with laparoscopic skills increased following the program, the faculty assessment of resident skills showed an upward trend but did not reach statistical significance.

At Yale University, within the Yale Laparoscopic Skills and Suturing Program, 291 trained surgeons performed 8730 standardized laparoscopic dexterity drills and 2910 intracorporeal suturing exercises.\textsuperscript{69} Ninety-nine residents performed the same drills and exercises the same number of times. Percentile graphs were prepared for each type of drill and suturing exercise and a database was developed for laparoscopic dexterity drills and suturing exercises, against which future trainees could assess their own level of achievement. In this study a significant correlation was noted between the time required for the suturing exercises and the time recorded in each of the dexterity drills (p 0.001). The authors concluded that this relationship strongly supports the correlation between bench model drills and a meaningful surgical task.

Derossis et al (1998) conducted a study to develop a series of structured bench model tasks designed to objectively measure laparoscopic skills.\textsuperscript{12} These tasks were used to evaluate the effects
of level of training and practice on performance. Forty-two subjects (residents and surgeons) viewed a 20 minute introductory video and then were tested on seven tasks such as peg transfer, mesh placement and endolooping, using a laparoscopic simulator. Performance was measured using a scoring system which rewarded precision and speed. Each candidate repeated all seven tasks and was rescored. Significant predictors of overall performance were (1) level of training ($p=.002$), (2) repetition ($p<0.0001$), (3) interaction between level of training and practice ($p=0.001$). The authors concluded that construct validity was demonstrated by measuring a significant improvement in performance with increasing residency level of training, and with practice.

### 3.2.4 The objective assessment of surgical skills in the human model

The Surgical Education Group at the University of Toronto considers bench model simulation with check lists and global rating forms to be a reasonable substitute for the operating room environment where the absence of standardization is an obstacle to reliable measurements of technical skills $^{62}$. 

On the other hand, while some aspects of technical skills, such as dexterity and eye-hand coordination, can be assessed by using bench models, the validity of simulators as surrogates for the assessment of surgical skills using real patients remains unproven. Real patients require more complex behaviours such as stress tolerance, judgement, and awareness of living tissue responses $^{60,25}$. Synthetic bench models may not represent anatomy well in their geometry or tactile properties and higher fidelity models are more expensive and less likely to be reusable $^{62}$. Even with animal models there are drawbacks (1) the animals are usually sacrificed following the training session and, therefore, the cost of error and the demand for excellence is reduced (2) there is an artificiality to animal sessions, which are conducted within a set time frame, unlike the human setting where the surgical procedure must run to completion regardless of its duration (3) in humans, the surgeon deals
with diseased and often distorted tissue, whereas animal models are generally healthy and, moreover, have their own unique organ structure and anatomic relationships (4) ethical considerations have placed serious constraints on the use of animals for surgical training and, in some jurisdictions such as the United Kingdom, the practice has been abandoned.

Rosser (1997) has stated that it is unclear whether laboratory drills and exercises can be used in isolation to credential surgeons. His position is supported by Cundiff who states that it is uncertain whether improvement in pelvic trainer skills translates to better operative skills in the human and, to some extent, this concern is supported by the fact that the faculty assessment of resident skills was not significantly improved following the Duke University laparoscopy training program. Derossis concluded that her bench model for laparoscopic skills training demonstrated validity by demonstrating a statistically significant improvement in performance with level of training and practice. However, she also stated that the model needed to be further evaluated by demonstrating a correlation with performance of analogous tasks using in vivo animal and human models.

A separate reason for wanting to assess surgical skills within the human pertains to the cost of administering examinations such as the OSATS on bench models. The examination format is labour-intensive and expensive. The OSATS developed by the Surgical Education Group at the University of Toronto cost approximately $200.00 (Cdn) per resident per iteration. This figure does not include the substantial amount of faculty time (at each administration) nor the funds that have gone into research and development.

In general surgery Winckel et al. (1994) developed a two part form for assessing the technical skills of general surgery residents in the operating room at the University of Toronto. The Structured Technical Skills Assessment Form (STSAF) consisted of two parts (1) a detailed
check list of approximately 120 items rated on a scale from 0 to 2 (2) an expanded summary scale in a global rating format to capture the assessment of the concept of technical competence, focusing on surgical behaviours instead of specific manoeuvres. Two faculty raters evaluated 12 residents who performed 41 operations. Interrater reliabilities were 0.78 and 0.73 for the check lists and global rating scales respectively with the two instruments being highly correlated (Pearson correlation coefficient=0.89). The authors concluded that check lists can provide highly specific feedback on a resident's technical skills; however, the global assessment takes less time to complete and can easily be applied to for each resident several times during a typical three or six month rotation. Kung (1993) used a global rating form, the Surgical Skills Form (SSF), similar to the one developed by Winckel to assess resident surgical skills in gynecology and obstetrics. One hundred and sixty-nine operations were performed at four university hospitals by residents from PGY1 TO PGY5. The average number of observations per resident was six. Interrater reliability for six raters was 0.65. Interitem reliability for 10 items was 0.90. The SSF appeared to be a reliable instrument for the assessment of surgical skills in obstetrics and gynecology and could reliably discriminate between residents at different levels (personal communication).

3.2.5 The objective assessment of laparoscopic skills in the human model

To date, no objective assessment of laparoscopic surgical skills in humans has been developed. Yet, the rapid expansion of laparoscopy in gynecology for a wide variety of conditions, the prolonged learning curve due to new instrumentation and skill sets, and the risks of surgical complications which may diminish the purported benefits, mandate the need for an assessment tool which is both reliable and valid. The assessment should be easy to administer and should be able to provide immediate feedback to residents, residency program directors, preceptors, clinical chiefs and hospital administrators. Ultimately, the assessment should ensure that individuals about to
embark in the practice of gynecologic laparoscopy have demonstrated adequate skills to perform procedures at a level of competence which meets national accreditation standards.

3.3 Summary

Technical skill is only one component of the overall construct of surgical competence, albeit an important one. During the past decade attempts have been made to objectify the assessment of technical skills so that scores assigned to examinees or trainees are value-free. The tools developed for measuring technical skills have been applied almost exclusively in bench simulation settings, using checklists and/or global rating formats. There are important reasons for wanting to assess technical skills within the human model, in particular because more complex behaviours such as stress tolerance, judgement, and awareness of living tissue responses can be measured. To date, no objective assessment specific of laparoscopic skills in the human model have been developed.
CHAPTER 4: THE LAPAROSCOPIC SKILLS INDEX (LSI)

4.1 Introduction

This thesis deals with the development and reliability testing of the LSI, a new measure of gynecologic laparoscopic surgical skills, developed by the author. It is designed to assess technical skills exhibited during level II procedures on patients at a single point in time and is therefore a "status" index. Potentially, the LSI could be used to monitor skills improvement and thus serve as a "change index". If shown to be reliable, valid, and easy to administer, the LSI could be a useful way of assessing technical skills in gynecologic laparoscopy within residency training programs and during preceptorships. Feedback would be immediate and ongoing, and could highlight specific areas which might require improvement.

A database, applicable to a multilevel spectrum of participants, could be developed against which residents and certified gynecologists would be able to measure their own level of performance. With time, the LSI could be modified and adapted to other surgical specialties.

Surgical competence is a multiattribute construct. The LSI is not intended as a measure of total surgical competence, but rather, a specific component within the larger concept.

4.2 Type of measurement

An issue of measurement development is whether checklists or global rating scales should be used. Checklists generally contain a list of detailed items of adequate performance and raters are asked to check off these items, using "yes/no" responses. Rating scales generally contain fewer items, are easier to administer, and focus more on broader issues of performance. In rating scales, a set of numerical values is assigned to responses that represent the degree to which a respondent possesses a particular attitude, value, or characteristic. The purpose of a rating scale is to distinguish among people who demonstrate different intensities of the characteristic being measured.
Several studies have shown that performance assessment by means of rating scales is as reliable, or even more reliable, than objectified checklists despite the fact that the former is considered to be more subjective and the latter more objective. The choice of rating scales or checklists in a test of performance may depend on the educational level of the examinees and the particular skill being tested. For isolated simple scales tested in lower classes of medical school or early residency training programmes, checklists may be preferred over ratings. However, checklists may reinforce rote memorizing strategies without an understanding of the outcome of a particular skill. For the same skill tested at postgraduate or postcertification levels, ratings may be preferred because less emphasis is placed on the details of a procedure, focusing more on effectiveness and the proficiency of the examinees' behaviour.

Thus, the LSI is designed as a rating scale because (1) global ratings are at least as reliable as checklists (2) rating scales are easier to administer than checklists (3) the measurement is targeted for postgraduate trainees and beyond.

Because the LSI measures a clinical activity it is classified as a clinimetric index. However, the author employed methods used for psychometric indices to evaluate reliability. (See 4.3.1.2, 4.3.1.3 and 5.1).

4.3 Scale construction

There are four main components to scale construction (1) item selection (2) selecting a method to scale responses (3) design (4) choosing a scoring method. Before entering into general usage the quality of the new measurement should be evaluated and modified as necessary through pilot studies and field trials that consider the following criteria (1) sensibility (2) reliability (3) construct validity (4) generalizability (5) responsiveness.

Feinstein stated that Sensibility is appraised with a "mixture of ordinary common sense
plus a reasonable knowledge of clinical reality \textsuperscript{72}. He divided sensibility into five major topics and several subdivisions. The main topics are (1) purpose and framework (2) overt format (3) face validity (4) content validity (5) ease of usage. \textit{Construct Validity} establishes the ability of an instrument to measure an abstract construct and the degree to which the instrument reflects the theoretical components of the construct \textsuperscript{73}. \textit{Generalizability} of a measurement justifies the inference of research outcomes to groups or situations other than those directly involved in the investigation\textsuperscript{73}. \textit{Responsiveness} is the ability of a measure to detect clinically important change, even if the change is small \textsuperscript{78}.

\textit{Reliability} refers to the consistency or reproducibility of the result given by an instrument measuring the same object repeatedly under constant conditions \textsuperscript{79}. This thesis deals with the construction and evaluation of the LSI from the standpoint of reliability. The other evaluative criteria are referred to again in \textbf{CHAPTER 10: DISCUSSION} of the manuscript.

\textbf{4.3.1 Item selection}

Indexes measuring level of achievement frequently are composed of multiple items. Because the multiple items are all intended to measure the same single attribute, psychometricians have given special attention to demonstrating that the many items are relatively "homogenous" \textsuperscript{80}. Single item indexes are rarely used in psychometrics because the selected phenomenon can seldom be adequately measured by a single expression \textsuperscript{80}. Individual items usually correlate poorly with the particular attribute. Also, each item has a degree of specificity in the sense of not correlating with any general attribute. Individual items have considerable random measurement error when rated on more than one occasion. With multiitem scales, the measurement error for individual items can be in either direction and tends to average out when item scores are summed to obtain a total score. As the number of items increases measurement error is reduced and reliability tends to increase \textsuperscript{81}.
While the LSI is a clinimetric index, it measures an attribute which is composed of multiple items.

Selecting items for a multiitem measurement, such as the LSI involves two processes. The first considers a pool of candidate variables (item generation) and the second selects the final items retained for the index (item reduction) 63,78.

4.3.1.1 Item generation

Item generation usually includes a review of items in similar scales if any exist, focus groups, clinical observation and the opinion of content experts 63,78,82. To ensure that a broad spectrum of questions are chosen for the initial pool of candidate variables, the investigator should have a hypothesis which describes the attribute’s domain of content including both the material to be tested and the population for whom the test is intended 81. Item generation ensures that all important variables are considered for inclusion in the scale and the initial item pool should yield a higher number of items than would be included in the final product 63,78.

4.3.1.2 Item reduction

Once the initial list of items is generated, the entire set should be checked to ensure that the scale has enough items and adequately covers the domain under investigation; in other words whether content validity exists 81. At this point a process of item reduction occurs. Item reduction reduces the index to a manageable number of items, eliminates inappropriate, redundant or ambiguous items, and selects variables which best describe the attribute of interest. The items selected should, at the very least, exhibit face validity, which represents a subjective affirmation by the experts or investigators that the items appear on the surface to be measuring the attribute of interest 72. Item reduction is often accomplished by consensus methods such as the Nominal group or Delphi techniques, opinions of the patients (subjects) on whom the index is to be applied, the
judgement of an experienced investigator and, finally, mathematical models which are used to perform item analysis after the pool of questions is administered to a "test" population. Item analysis furnishes a variety of statistical data regarding how subjects responded to each item and how each item relates to overall performance. The best items on any test are the most discriminating. They are less ambiguous, cannot be of extreme difficulty and tend to make individual differences on the final test more reliable. The simplest discrimination index is the ordinary item-total correlation ($r$) between each item and the total test score. An improved version of the item-total correlation is the corrected item-total correlation ($r_c$) in which the specific item being evaluated is eliminated from the total score so as to avoid a spuriously inflated result which would be based in part on the item's correlation with itself. Items that fail to correlate with the total score are generally either excessively difficult or easy, ambiguous, or unrelated to the attribute of interest. A cutoff of 0.30 for the corrected item-total correlation is an arbitrary guide to defining a discriminating item.

Item analysis programs may also include empirical estimates of trace lines referred to as item characteristic curves. Two basic properties of a trace line are its difficulty and its discrimination. Difficulty refers to how much of the attribute an individual must possess to achieve a given level of response (performance) while discrimination describes the extent to which the probability of a particular level of response (performance) correlates with the attribute. An item with a perfectly flat trace line does not discriminate and should be eliminated from the test. Most models are called monotonic in that the probability of a specific response (denoted as alpha) is expected to increase with the attribute in the general form of an ogive. Making an item more discriminatory increases the slope portion of the ogive.

A commonly used term, frequency of endorsement, refers to the proportion of people ($p$) who
give each response to an item. It is a function of the item difficulty. Items with extremely high or low endorsement rates ($p > 0.95$ or $p < 0.05$) are eliminated because they add little to the discriminatory power of the index and make the test unnecessarily longer. In practice, only items with endorsement rates between 0.20 and 0.80 are used.

### 4.3.1.3 Item homogeneity

In measuring an attribute with a multiitem index it is essential that each item correlates with the total score and, moreover, the items should be moderately correlated with each other. These two factors form the basis of the various tests of homogeneity or internal consistency of a scale. The corrected item-total correlation has been discussed above. Summary statistics, called Cronbach's coefficient alpha for items with ordinal responses, and Kuder-Richardson 20 for items with dichotomous responses, define the internal reliability of an index. They are obtained from the average intercorrelation among all items. A high value for these intercorrelation statistics helps confirm the underlying assumption that the component items are related to the single attribute of interest. The minimal satisfactory level of homogeneity for intercorrelation is 0.80 for group research concerned with mean differences among experimental treatments and 0.90 for decisions pertaining to individual subjects. Two problems with Cronbach’s alpha are (1) it is sensitive to the number of items as well as the magnitude of the intercorrelations. Thus, a scale can be made to look more homogeneous simply by doubling the number of items (2) if alpha is too high it may suggest some level of item redundancy, that is, a number of items asking the same question in slightly different ways.
4.3.2 Selection of a method for scaling responses

4.3.2.1 Categorical versus continuous response options

Categorical judgements limit the choice of response levels. For example, a dichotomously scored item (pass-fail, present-absent, good-bad) can distinguish between only two levels of an attribute. Most measurement problems require much finer differentiation. Thus, categorical judgements should not be employed in circumstances where the response may be on a continuum. This leads to uncertainty around what constitutes a positive response and a potential loss of information with a corresponding reduction in reliability. In addition, using a categorical judgement where a continuum might be appropriate leads to decreased efficiency of the instrument and, consequently, a greater sample size necessary to show an effect.

For continuous judgements there are three broad category approaches to scaling (1) direct estimation techniques such as visual analog scales (VAS) and adjectival scales such as the Likert scale in which subjects are required to indicate their response by a mark on a line or a check in a box (2) comparative methods such as the Thurstone and Guttman techniques in which subjects choose among a series of alternatives which have been previously calibrated by a separate criterion group (3) econometric methods such as the standard gamble and the time trade off techniques. Econometric methods were devised to assign numerical values to various health states and are frequently used in determining cost/benefit ratios. They are described elsewhere and are not appropriate to measuring laparoscopic skills. They will not be discussed further.

While VAS techniques have been used extensively in medicine because of their simplicity, they are one item tests and therefore likely to have lower reliability than multiitem scales. Moreover, the attribute of interest, for example functional capacity, is often complex and cannot be defined by a single item. One would need multiple VAS's to assess related aspects of the attribute.
The Likert (summative), Thurstone (differential) and Guttman (cumulative) scales are adjectival and contain descriptions with either discrete or continuous responses. Each is a multiitem index that measures a single attribute, but the methods work differently.

Thurstone scales require that a subject's response specify where he or she falls along some attitudinal dimension in which the responses have previously been ranked by a panel of "expert" judges. This method is not used much today. Scale construction is time-consuming and, further, the scales quickly become outdated if there are changes in society affecting the attitudes being measured. Also, the questions in a Thurstone scale are double-barreled and confusing to respondents. Responses at the far end of the spectrum are blurred, creating difficulty in assigning a scale score.

Guttman scales tend to position individuals along some attitude continuum by assigning a single scale score. The items are sequentially related to each other and only one unique combination of responses can achieve a particular score. The technique is appealing for clinical situations where it would be interesting to document hierarchical scales or behaviours such as a sequence of ambulation, where each level achieved subsumes all those below it. Guttman scales are no longer frequently used. For one thing, most psychometric and clinical phenomena do not follow an exact hierarchical pattern. Second, Guttman scales can achieve a perfect, error free pattern using a series of items which are not related to a common attribute. Third, the final index has an ordinal rather than a dimensional scale and, finally, discrimination along the attribute of interest is usually coarse because only a limited number of questions will fit the required staircase pattern.

In the Likert method, the numerical values assigned to the response categories for each question are simply combined to produce a final index score. Likert index items have a sigmoidal
shaped item characteristic curve or "ogive", rising gradually with a steep central slope and then flattening out. The three underlying assumptions of a Likert index are (1) that the trace line for each item forms a characteristic monotonic curved line (2) that the sum of the family of item characteristic curves yields a roughly linear relationship between the score and level of the attribute (3) that all questions tend to measure the same single phenomenon or attribute. The last of these assumptions has three important applications for Likert indexes. First, indexes intended to measure the same attribute may have different component items, but if all components of an index measure the same attribute, different arrangements of the items should yield similar results. Second, adding more questions or items to an index should improve its ability to discern among individuals. Third, because all components are measuring the same attribute, differential emphasis or weighting of the items should not affect the final result. Likert scales are one of the most widely used methods of scaling in the social sciences today. They are easier to construct and tend to be more reliable than other scales with the same number of items. In constructing a Likert scale it is critical that the test population be carefully chosen to represent the people for whom the index is intended and it must be tested under conditions similar to the index's ultimate usage. One problem of Likert scales relating to reproducibility is the fact that there are many ways of achieving the same scale score. If a respondent scores high on several items, but low on several others, the summated score would reflect a moderate position. A different respondent could be assigned the same scale score even though he or she had neither positive nor negative responses to any of the items.

4.3.2.2 Scaling response method of the LSI

The LSI uses the Likert model as the method for scaling responses. It is a multiitem, adjectival, global rating scale in which each item is scored on a continuum.
headings (1) item generation (2) initial item reduction (3) the LSI design (4) the LSI scoring method (5) methods of administering the LSI.

4.4 Construction of the LSI

4.4.1 Item generation for the LSI

Item generation began in January, 1996. Initially, the author conducted a Medline search for publications pertaining to surgical skills assessment dating back to 1970 using the following mesh headings, key words and combinations: surgery, laparoscopic surgery, measurement, index, skills assessment, rating scales, and several subheadings. Few publications were found and they are described in CHAPTER 3: MEASURING LAPAROSCOPIC SURGICAL SKILLS. There were no publications pertaining to the objective assessment of laparoscopic skills. Since the work on this thesis began a number of other publications have appeared in the literature, including assessments of laparoscopic skills using bench model simulations. These are also described in CHAPTER 3. To date, no articles dealing specifically with the objective assessment of laparoscopic skills in humans have been published.

A modification of the Delphi Consensus method was then used to generate the initial item list. The Delphi method attempts to obtain expert opinion in a systematic manner. Experts who participate are polled individually and anonymously with self-administered questionnaires. The survey is conducted over three or four "rounds" but after each one, the results are elicited, tabulated and then reported to the group. A Delphi is considered complete when there is a convergence of opinions or when a point of diminishing returns is reached. There are limitations to the Delphi method including the need for a sizeable group of experts, panel fatigue after several rounds, and lack of personal contact among participants. The process used by the author did not strictly adhere to the anonymity of the Delphi process because it was conducted in three phases, of which the
second was neither anonymous nor did it include the entire group of content experts.

In January, 1995 the author sent a letter to 14 colleagues in University of Toronto teaching hospitals and Ontario community hospitals experienced in gynecologic operative laparoscopy (Table 1). All 14 were members of the American Association of Gynecologic Laparoscopists. In addition to their certification in gynecology, each obtained extended training in operative laparoscopy through preceptorships or accredited "hands on" laparoscopic skills training programmes. Two of the experts have pioneered techniques in operative laparoscopy in the areas of urogynecology and multiorgan endometriosis, respectively. These two individuals also received the University of Toronto award for continuing education having developed courses in operative laparoscopy which provided bench, animal, and human laparoscopic experience for registrants. Two other content experts participated as faculty in these courses. The remaining individuals were experienced laparoscopy teachers, including the community hospital experts each of whom offered training electives in gynecologic laparoscopy. The reasons for choosing community experts as well as University of Toronto experts were two fold (1) to broaden the scope of opinion regarding the components of laparoscopic skills, (2) to create an index that would be relevant for credentialing individuals applying for hospital privileges in gynecologic laparoscopy.
Table 1: Background Information on Content Experts

<table>
<thead>
<tr>
<th>Expert</th>
<th>Gynecology Practice</th>
<th>Practice Location</th>
<th>Operative Laparoscopy Experience (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>complex endometriosis</td>
<td>U of Toronto</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>urogynecology</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>generalist</td>
<td>community, GTA</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>generalist</td>
<td>community, GTA</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>generalist</td>
<td>community, Orillia</td>
<td>5</td>
</tr>
<tr>
<td>14*</td>
<td>generalist</td>
<td>U of Toronto</td>
<td>5</td>
</tr>
</tbody>
</table>

* did not respond to letter requesting items for LSI

These 14 content experts were apprised of the author's intention to develop a laparoscopic skills index and were asked to identify what they considered to be essential basic skills in gynecologic laparoscopy. To facilitate their responses, the author provided an example of a basic skill, a skill subheading, and checklist items which might be incorporated in adjectival descriptions of a skill. The examples were as follows:

"A basic skill might be laparoscopic knot tying. A subcategory would be Endoloop ligaturing. Checklist items would include (1) appropriate loading of the Endoloop, (2) application of the loop around the pedicle without incorporating other structures, (3) tightening of the knot using a double cinching technique".
Thirteen of the 14 mail recipients responded by March, 1995. The unabridged list consisted of 178 candidate variables.

4.4.2 Initial item reduction

Of the 178 variables listed an initial item list was developed as follows:

1. elimination of duplications (119, residual 59).
2. elimination of "procedure" variables. This refers to the naming of total procedures, such as oophorectomy or removal of ectopic pregnancy, as opposed to specific laparoscopic skills (total 8, residual 51).
3. elimination of items which are not technical skills, for example, positioning the patient, preparing and draping the patient (total 4, residual 47).
4. elimination of subcategories. This refers to variables which are subsumed by a broader definition of a skill. For example, Endoloop ligaturing is a subcategory of knot tying, and inspection of the upper abdomen is a subcategory of initial inspection (total 12, residual 35).
5. elimination of descriptor/checklist variables. For example, a checklist item pertaining to the skill of Veress needle insertion would include "chooses target just below sacral promontory". A checklist item for the anatomic inspection of the pelvis would include "exposes and assesses the anterior and posterior aspects of each ovary" (total 11, residual 14).
6. elimination of infrequently named variables (≤3 times) such as use of the ultrasound scalpel (total 4, residual 10).

Thus, the initial item list consisted of 10 items as follows.

1. Veress needle insertion/trocar insertion
2. initial inspection
3. laparoscope control
4. suction/irrigation
5. dissection
   i) laser dissection
Following construction of the initial 10 item list, citywide gynecologic laparoscopy rounds were conducted in May, 1995. Seven of the original panel of content experts attended and each was asked to review the initial item list. All agreed that the first portion of item 7, suturing, should be removed because it is rarely used in gynecologic laparoscopy, with the exception of urogynecology and it would not be a prerequisite for proficiency in level II laparoscopic procedures required for specialty certification. One item, instrument knowledge, was added to the list. While not strictly constituting a technical skill, the panel members felt that there were several aspects of instrument knowledge, specific to laparoscopic surgery, which would be important in mastering the various skills. These include laser and electrocautery settings, choice of trocars, and types of morcellators.

Thus, a second item list was created which included a modification of one item and the addition of another for a total of 11 items:

1. Veress needle insertion/trocar insertion
2. Initial inspection
3. Laparoscope control
4. Suction/irrigation
5. Dissection
   i) Laser dissection
   ii) Electrocautery dissection
   iii) Blunt dissection
6. Hemostasis
7. Ligation
8. Stapling/clipping
9. Tissue removal
10. Final inspection
11. Instrument knowledge
For each item, an adjectival descriptor was created by the author, drawing upon many of the subcategories and checklist items contained within the initial unabridged pool of variables. The 11 item list, along with descriptors, was then sent to each of the original content experts to solicit opinions regarding the comprehensiveness of the item list for the application intended and for any suggested additions, deletions or modifications of the items and/or descriptors.

All content experts agreed with the 11 item list. Only one person made suggestions for modifications of the adjectival descriptors, with the intent of reducing the detail to a more global type of rating. Because all but one individual accepted the descriptors in the format presented to them, the modifications suggested by this individual were maintained in the author’s study file but not applied at this point in time.

4.4.3 The LSI design

As mentioned previously, the scaling technique chosen for the LSI is a modification of the Likert model. This technique was chosen because Likert models are easy to construct, tend to be more reliable than other scales with the same number of items, are used widely in psychometric scaling and have an underlying assumption of ogive trace lines which is suitable for most stimuli.

The LSI was designed with five categories of response for each item. This is the same as the surgical skills form and global rating form used for laparotomic surgery in gynecology and general surgery, respectively. The minimum number of response categories for a Likert scale should be in the region of five to seven. Studies have shown that for reliability coefficients in the range normally encountered, from 0.4 to 0.9, the reliability drops as fewer categories are used. Going from 10 categories to five reduces the reliability by about 12 percent, often because extreme positions on the scale are not used. However, when more than five categories are used it becomes difficult to meaningfully describe the categories in English and distinctions may become too fine.
Practically, using five categories, ranging from poor to excellent, seemed sufficient for the intended purpose of the index. For example, there would be very little advantage to differentiating poor skills from very poor skills, or excellent from outstanding. This was agreed upon by the experts.

Initially, the LSI rater package consisted of three pages (1) rater instruction and scoring guidelines, Appendix 4 (2) item descriptors, Appendix 5 (3) rater scoring form, Appendix 6. The scoring form response categories were ranked numerically from 1 to 5 with a separate category, number 6, reserved for not observed items. For example, tissue removal might not be necessary in all laparoscopic procedures. A score of 1 was classified as poor and a score of 5 as excellent.

This initial design layout for the LSI was used during a pilot reliability study. A second reliability study which is the central theme of this thesis employed a revised version of the LSI (see CHAPTER 8: MEASURES TAKEN TO IMPROVE THE DESIGN, RATER PREPARATION AND CONTENTS OF THE LSI).

4.4 The LSI scoring method

As with other Likert-type scales, the measured level of the LSI is the sum of the scores obtained by the individual items. This is the standard method for scoring equally weighted linear combinations of item responses. The evidence that differential weights seldom make a difference is substantial. If the scale is comprised of relatively homogeneous items, where regression weights will also be within a fairly narrow range, the effect of weighting is minimal. Differential weighting is sometimes of benefit when (1) the number of items is relatively small (<20) and (2) item-total correlations vary markedly. Differential weightings might increase the reliability slightly in these circumstances, but the same increase in reliability might also result from adding two or three new items.
Because some items in the LSI would be observed on one occasion and not on others, the sum score is a crude score. The final score is expressed as a percentage derived by the formula:

\[
\left[ \frac{\sum_{i=1}^{n} x_i}{n} - 1 \right] \times 25
\]

where \( x_i \) are the individual items collected, ranging from 1 to 5 and \( n \) is the total number of items. An average is taken of all the item scores as they were collected (range 1 to 5) and then one is subtracted from this average giving a range of 0 to 4 which is multiplied by 25 to obtain the final result (range 0 to 100).

4.4.5 Methods of administering the LSI

1. In vivo assessments

Given the brevity of the LSI it should be possible for trained raters to complete an assessment in less than five minutes. In postgraduate training programs the faculty raters would complete the assessment using rater forms kept at the main nursing station or within the gynecology operating theatre. The completed assessment form would be forwarded to the postgraduate training coordinator. Each resident would be rated several times during any three or six month rotation on the gynecology service. On completing the rotation he/she would be given a final evaluation based on the individual assessments. The final evaluation would be incorporated into the overall ITER. This method of administration is similar to that currently used within the University of Toronto Department of Obstetrics and Gynaecology for clinical clerks. After each encounter with a clinical
clerk (on wards, clinics, and in operating theatres) the faculty member completes an evaluation form and the evaluations are aggregated to create a final mark.

2. Videotape assessments

Because laparoscopic surgery is performed by observing the abdominal contents on a video monitor, it is common practice to videotape the operative procedure. This is done to maintain a permanent record for review with patients, for medical legal usage and for creating a library of educational videotapes.

LSI ratings of videotaped procedures could have several advantages:

1. a rater could assess the trainee’s skill at any suitable time, even using a VCR at home.
2. with a videotaped record the rater would not have to rely on memory to recall the events of the procedure.
3. a blinded rater could be used who did not participate in the surgery, because videotaping operative laparoscopy captures only the intra-abdominal events and it is impossible to identify the operator.
4. if a rater is unsure about scoring a skill on the initial viewing, he/she could take advantage of the rewind and slow motion features of videotaping.
5. the videotaped procedures can be used as educational tools to illustrate how a skill should be performed, as well as the levels of performance of said skill.
6. videotapes may be used as a mechanism for training evaluators, for example, in establishing what constitutes a low score versus a high score.

There may be disadvantages to videotaping as well:

1. if the procedure is edited rather than viewed in total, then the editing process may be subjective and, therefore, certain skills might be insufficiently represented or the chosen segments
might be biased in either direction.

(2) a videorecording makes it impossible to tell which part of the procedure is done by the surgeon being assessed versus the portion done by the surgeon not being assessed.

(3) videotaped reviews require off hours viewing rather than on the spot viewing and therefore entail a sacrifice on the part of the faculty rater.

(4) certain items are not video recordable. For example, instrument knowledge, is not visible nor is the insertion of the Veress needle. With respect to the latter, trocar insertion could be used as a surrogate for the Veress needle insertion. Certain aspects of instrument knowledge, such as the choice of instrument and how it is used, could be captured within the assessment of several other items (see descriptors for laparoscope control, suction, irrigation, tissue removal, laser dissection, etc).

(5) whereas in vivo scoring permits immediate feedback, videotaped scoring could result in a delay in feedback which would reduce the amount of time available during any one rotation for specific remedial work.

4.5 Summary

A new index has been developed to measure laparoscopic skills in gynecology, the Laparoscopic Skills Index (LSI). A cadre of content experts was used to generate the pool of variables and the initial item list. Based on a set of rules applied to the 178 items of the initial list, a final list of 11 items remained to define the pilot version of the LSI. The scaling method uses a Likert (summative) model with a five category, global rating technique of response. Descriptors were developed for the individual items. The two methods of administering the LSI are (1) in vivo (2) by videotape. The pros and cons of each method are discussed.
CHAPTER 5: MEASUREMENT RELIABILITY

5.1 Definition of reliability

Reliability refers to the consistency or reproducibility of the result given by an instrument measuring the same object repeatedly under constant conditions 79.

This thesis is concerned with the assessment of the reliability of the LSI. Demonstration of reliability is a prerequisite to establishing the validity of an instrument. The upper limit of an instrument's validity is a function of its reliability 95.

Two main traditions exist for the assessment of the reliability of continuous measures 79. First, the psychometric tradition was developed in the context of questionnaires and composite indexes. Cronbach proposed the coefficient alpha to estimate the reliability of a summation of items. Because the items should show high intercorrelations to capture the concept of interest, the use of internal consistency to label this form of reliability is intuitively sound. Second, the analysis of variance (ANOVA) tradition was developed in the context of the assessment of rater performance, both in terms of intrarater and interrater reliability. Because it estimates the equivalence of repeated measurements made on the same subjects, this form of reliability assessment is named reproducibility. Internal consistency was discussed in 4.3.1.3 This chapter focuses on reproducibility.

5.2 Measurement error

Measurements are rarely perfectly reliable. All instruments are fallible to some extent and all humans respond with some inconsistency. For any observed score ($X$) there is a true score ($T$) and an error component ($E$) which is random and normally distributed with zero mean and a variance $E^2$. The expression implies that for any given measurement ($X$), a hypothetically true or fixed value exists ($T$), from which the observed score will differ by some unknown amount ($E$). The
true component is the score the subject would have gotten had the measurement been taken by a perfect measuring instrument under ideally controlled conditions. The difference between the true value and the observed value is measurement error. Two basic types of measurement error exist. Systematic errors are predictable errors of measurement. They occur in one direction, consistently over estimating or under estimating the true score. Such error is also called bias. If detected, it is a matter of recalibrating the system or adjusting for the error by adding or subtracting the appropriate constant. Random errors are due to chance and can affect any single subject’s score in an unpredictable way from trial to trial. For example, if a patient moves slightly or does not stand fully erect each time his height is measured, the measurements will be inconsistent.

Reliability focuses on the degree of error present within a measurement system. As random errors diminish, the observed score moves closer to the true score and the measurement is more reliable.  

5.2.1 Sources of measurement error

Measurement errors can be attributed to the individual taking the measurement (rater, observer, clinician, etc.), the measuring instrument or procedure, or the variable being measured (subject, patient). Measurement error can be reduced by efforts to reduce the variability arising from the various sources. Often, this is accomplished during “tryouts” or pilot studies.

5.3 Rater reliability

All clinical measurements require that a human observer or rater is part of the measurement system. In most situations, the rater must apply operational criteria to subjective observations, as in the grading of the symptoms of pelvic endometriosis. Test-retest reliability is used to establish that an instrument is capable of measuring a variable with consistency. In a test-retest study, the sample of individuals is subjected to the identical test on two separate occasions, keeping all test
conditions as constant as possible, such as tester, time of day, and environment. The correlation coefficient derived from this type of analysis is called a test-retest reliability coefficient. Rater reliability is actually a special case of test-retest reliability where the instrument and response variables are considered stable, so that any differences between scores are attributed to rater error. Intrarater reliability refers to the stability of data recorded by one individual across two or more trials. When carryover or practice effects are not an issue, intrarater reliability is usually assessed using trials that follow each other within short intervals. Interrater reliability concerns variation between two or more raters who measure the same group of subjects. Even with detailed operational definitions and equal skill, different raters are not always in agreement about the quality or quantity of the variable being assessed. Interrater reliability is best assessed when all raters are able to measure a response during a single trial where they can observe the subjects simultaneously and independently. This eliminates true differences in scores as a source of measurement error when comparing raters’ scores. Videotapes of patients performing activities have proved useful for allowing multiple raters to observe the exact same performance (see Overview of Research Methods, 6.3).

5.4 Measuring reliability

Reliability estimates are based on the statistical concept of variance, which is a measure of the variability or differences among scores within a sample. The larger the variance the greater the dispersion of scores; the smaller the variance, the more homogeneous the scores. Reliability measurements estimate the extent to which the test score is free from error. That is, reliability is a measure of how much of the total variance is attributable to true differences between scores. Thus, reliability can be expressed as a ratio of the true score variance to the total variance, or $T / (T+E)$. This ratio yields a value called the reliability coefficient. In statistical terminology the relationship
is expressed as

\[ r_{XX} = S_X^2 - S_E^2 / S_X^2 = S_T^2 / S_X^2 \]

where \( r_{XX} \) is the symbol for a reliability coefficient. \( S_X^2 \) represents variance of the observed score, \( S_E^2 \) represents the error variance and \( S_T^2 \) represents variance of the true score.

5.4.1 Correlation and agreement

Many reliability coefficients are based on measures of correlation. Correlation reflects the degree of association between two sets of data or the consistency of position within the two distributions. In other words, correlation reflects covariance or rank order characteristics within the data. Commonly used correlation coefficients are the Pearson Product Moment Correlation Coefficient (for interval/ratio data) and the Spearman Rank Correlation Coefficient (for ordinal data). There are several problems using correlation as the sole measure for reliability. First, correlation does not express agreement. Yet, reliability is concerned with the degree of similarity (concordance) between repeated scores, not just the consistency of their ranks. Second, the Product Moment Correlation Coefficient is designed to assess only bivariate relationships. That is, only two ratings or raters can be correlated at one time. Third, correlation is not useful for assessing the different aspects of reliability, such as raters, test forms and testing occasions, and for separating out variance components due to error or true differences in the data sets. Thus, the correlation coefficient does not evaluate the full scope of reliability and is not a true reliability coefficient.

To overcome the limitations of correlation as a measure of reliability, researchers sometimes use more than one reliability index within a single study. For example, both correlation and a t-test can be performed to assess consistency and average agreement between data sets in a test-retest situation. However, a better approach would be to have a single index that reflects both the degree of correspondence and agreement among ratings. Such an index is provided by the intraclass
The ICC is a reliability coefficient which combines a measure of correlation with a test of the difference in means. In “regression language”, the ICC assesses not only the similarity of slopes, but also the similarity of intercepts. The ICC, unlike the Pearson Correlation statistic is sensitive to systematic bias. Thus, if one rating is systematically higher or lower than the other, the ICC will be correspondingly reduced.

The ICC derives its mathematical definition from a repeated measures ANOVA. The total variance among the various measurements or judgements is apportioned among three sources: the differences among methods (or observers), the differences among subjects, and a remaining “unexplained” residual error variance.

There are several advantages to the ICC (1) it can be used to assess reliability among two or more ratings (2) it does not require the same number of raters for each subject, allowing for flexibility in clinical studies (3) it is designed primarily for use with interval/ratio data, yet can be applied without distortion to data on the ordinal scale, when intervals between such measurements are assumed to be equivalent. Under most circumstances, unless the distribution of scores is severely skewed, one can analyze the data from category rating scales as if they were interval, without producing severe bias. (4) the ICC supports the model of generalizability which is based on the idea that differences between observed scores are due to a variety of factors, not just true score variance and random error. In generalizability theory, error variance is multivariate and is further partitioned to account for the influence of specific facets such as testing conditions, time of day and observers. The generalizability coefficient is an extension of the reliability coefficient:

\[
\text{ICC} = \frac{S^2_T}{S^2_T + S^2_F + S^2_E}
\]

where \(S^2_T\) and \(S^2_E\) are the variances in true scores and error components, and \(S^2_F\) is the variance of the facets of interest.
Shrout and Fleiss describe three models of the ICC according to how raters are chosen and assigned to subjects during a reliability study\textsuperscript{102}. Each ICC model can be expressed in two forms, depending on whether the scores are single ratings or mean ratings. Most often, reliability studies are based on comparison of scores from individual raters (or ratings). There are times when the mean of several raters (or ratings) may be used as the unit of reliability. For instance, when measurements are unstable, it may be necessary to take the mean of several measurements as the individual score to obtain satisfactory reliability. In this situation, mean scores have the effect of increasing reliability estimates as means are considered better estimates of true scores, theoretically reducing error variance.

5.5 **Interpretation of reliability coefficients**

ICC's should range from 0.00 to 1.00 with values closer to 1.00 representing higher reliability. However, it should be noted that a negative value for ICC can occur when there is a greater variability within the object being measured than between objects. When this occurs one may conclude that reliability does not exist for the particular measure. While judgements are best made within the context of each study, general guidelines suggested by Portney and Watkins are as follows (1) below 0.50 represents poor reliability (2) from 0.50 to 0.75 is moderate reliability and (3) above 0.75 is good to excellent reliability \textsuperscript{73}. Rosner\textsuperscript{103} recommends the following guideline (1) below 0.40 poor (2) 0.40 to 0.75 fair to good (3) equal to or greater than 0.75 excellent.

5.6 **Summary**

The reliability of a measurement tool requires demonstration of both internal consistency and external reproducibility (rater reliability). An inverse relationship exists between measurement reliability and random error. Random error can be diminished by reducing the variability arising from various sources including the rater, the procedure, or the variable being measured. In
measuring rater reliability, we are concerned not just with the consistency of ranking between raters. but with the level of similarity (concordance) between scores. Thus, correlation as a measure of reliability is limited. Indexes such as the Intraclass Correlation Coefficient (ICC) which reflect both the degree of correspondence and agreement among ratings are preferred. Several models of the ICC exist according to how raters are chosen and assigned to subjects during a reliability study. Numerical guidelines defining levels of reliability were presented.
CHAPTER 6: RELIABILITY TESTING OF THE LSI

This thesis deals with two aspects of the LSI (1) construction of the LSI (2) reliability testing of the LSI. Reliability is one of the criteria listed in 4.3 which defines the quality of a new measurement tool. Chapters 6, 7, 8 and 9 are concerned with the reliability testing of the LSI.

6.1 Research objective

To evaluate the reliability of a newly developed measure of gynecologic laparoscopic surgical skill known as the Laparoscopic Skills Index (LSI).

6.2 Research questions

6.2.1 Primary research question

1. To what extent do the items comprising the LSI correlate with each other and the total score achieved by summation of the individual item scores? That is, what is the internal consistency (homogeneity) of the LSI?

2. What are the intrarater and interrater reliabilities of the LSI? That is, to what extent does the LSI exhibit similar results (reproducibility) when repeatedly applied to the same subject under constant conditions?

6.2.2 Secondary questions

3. Is the level of internal consistency sufficient to describe the LSI as a unidimensional index?

4. Is the level of interrater reliability sufficiently high to proceed to a study of construct validity of the LSI as a means of objectively assessing laparoscopic skills among gynecologists?
6.3 Overview of research methods

In this study videorecordings of laparoscopic surgery performed by gynecologic residents and faculty were obtained. The use of videotapes allowed for multiple raters to observe the exact same performance on each subject. The study was conducted in two phases (1) a pilot study in which six raters each reviewed six videotaped procedures (subjects), on two occasions, separated by one month (2) a main study in which four raters each reviewed 20 videotaped procedures (subjects).
CHAPTER 7: PILOT LSI RELIABILITY STUDY ("PILOT STUDY")

The pilot reliability study was conducted prior to the main study to determine how well the process would be received by raters and subjects and to determine what modifications would be necessary to improve the contents, instructions, criteria, or other arrangements of the index’s construction before conducting the main reliability study.

7.1 Pilot study summary of methods

The pilot study was conducted in 1998. Videotaped records of six gynecologic laparoscopy procedures were edited to approximately 30 minutes in length. Each videotaped procedure contained representative segments of the videorecordable LSI items. The videotaped procedures were compiled onto one videotape. Six raters reviewed each compilation on two separate occasions. The order of the procedures on the first and second compilations differed.

7.2 Pilot study subject (procedure) selection

All procedures were performed by postgraduate residents in gynecology at Women’s College Hospital under the supervision of a faculty member. In three of the procedures the postgraduate trainee was in the final precertification year (PGY5). In one procedure the trainee was at the PGY4 level, and in the remaining two, the PGY3 level.

7.2.1 Pilot study, procedure eligibility criteria

Each laparoscopic procedure used for the pilot study met the following eligibility criteria:

1. the surgery was booked for no less than 60 minutes
2. the procedure was booked in a designated gynecologic laparoscopy operating room
3. the procedure was performed on an elective basis during daytime operating hours
4. all operations qualified as level II laparoscopic procedures in accordance with the guidelines for certification in laparoscopic procedures prepared by the Society of
7.2.2 Pilot study procedure exclusion criteria

The following were excluded from the study:

1. nonelective (emergency procedures)
2. level I procedures, such as diagnostic laparoscopy and laparoscopic sterilization, and level III procedures, such as the Burch sling for urinary stress incontinence
3. procedures where the supervising faculty member was one of the pilot study raters.

7.3 Pilot study rater selection

Of the six raters, four were selected from the initial cadre of 13 content experts. The other two were faculty gynecologists, one from Women’s College Hospital and one from the Toronto Hospital. In total, the rater distribution consisted of three raters from Women’s College Hospital and one each from The Toronto Hospital, Mount Sinai Hospital and the Toronto East General Hospital. The raters were “randomly” chosen and, therefore, the results obtained could potentially be generalized to other raters with similar characteristics. This randomness was, of course, theoretical in that we do not have access to the entire population of potential raters. The six raters were identified alphabetically, A to F.

7.4 Pilot study ethics approval

The study was submitted to the Women’s College Hospital Research Ethics Committee for approval. Information sheets and consent forms were developed for the participating faculty surgeons and postgraduate trainees (Appendix 3). The procedure videotapings were for the determination of LSI rating scores only and were not to be used in any manner to guide treatment or alter management of the involved patients. The LSI scores obtained were not to be used in the trainees’ end-of-rotation ITER reports. The videotapes were anonymous, with no notation of the
patients’ or surgeon’s names or the dates of surgery. All tapes were returned to the supplying surgeons following completion of the study. The disposition of the tapes were at the surgeons’ discretion. The study was approved by the Research Ethics Committee.

7.5 **Pilot study videorecordings and videotape compilations**

The six laparoscopic procedures fulfilling the above criteria were selected for videotaping on a first come- first serve basis following approval of the Research Ethics Committee. The operative procedures were performed in the usual fashion with no modifications for the purpose of videorecording. All procedures were performed in one of two dedicated gynecologic laparoscopy suites at Women’s College Hospital. Videotape recordings were made from the video monitor images captured by a camera attached to a 10 mm operating laparoscope. The videorecordings were begun at the time the laparoscope was introduced into the abdominal cavity and were continued uninterrupted for the duration of the procedures.

Videorecordings were obtained using high quality super-VHS tapes (TDKXP superpro/ST-120, Fuji S-VHS/ST 160) and S-VHS recorders (Sony). The videotaped procedures were then edited for length to approximately 30 minutes each, such that representative segments of each videorecordable LSI item were included. The six operative procedures were compiled onto a single tape with titles that clearly defined when one operative procedure ended and the next began. The operative procedures were arranged initially in chronologic order, however, the order seen in the final compilation was determined by computerized random number generation. A second compilation of the procedures was then made after rearranging the order of the procedures, again using random number generation.

All videorecordings and video editing were performed by the author’s coinvestigator (RP).
7.6 Pilot study rating procedure

Each rater who consented to participate in the study was supplied with the following:

1. a rater instruction sheet (Appendix 4)
2. six LSI rating forms, each consisting of two pages, with item descriptors on page 1 and a scoring form on page 2 (Appendix 5, Appendix 6)
3. a videotape compilation.

The raters were instructed to complete the scoring of a procedure as soon as the viewing of that procedure was complete. While the LSI item list contained 11 items (see Initial item reduction 4.4.2), the raters were given a shortened version in which items 7, 8, and 11 were omitted. The reasons for these omissions were (1) none of the procedures that were recorded contained items 7 and 8 (2) number 11, instrument knowledge, is not a videorecordable entity.

One month after the initial ratings were completed the raters were provided with the second compilation of the previously viewed procedures arranged in a new order. The rating procedures were repeated.

The initial and repeat viewings were labeled Time 1 and Time 2, respectively.

The data sheets were collected and entered into a Microsoft Excel 97 worksheet by one of the investigators (RP). A second check of the data for entry errors was conducted by the author (JS).

7.7 Pilot study outcomes

7.7.1 Internal consistency

The homogeneity of the items was assessed by determining item intercorrelations and the correlations of the individual items with the total score.

7.7.2 Rater reliability

The total scores assigned by each rater for each operative procedure and each viewing were
calculated. The intrarater and interrater variabilities were then calculated. Intrarater variability was based on 36 paired ratings (six raters reviewing six operative procedures on two occasions). Intrarater variability was defined as the variation in scores for each rater on two occasions. All differences were taken as the absolute value of the first rating minus the second rating. Intrarater variability was based on 180 observer pairings (six operative procedures viewed twice by 15 rater pairs). Interrater variability was defined as the variation in scores among the six raters.

7.7.3 Rater comments

A space was provided on the rating form for written comments and raters were asked to comment on any difficulties encountered while scoring a segment.

7.8 Pilot study statistical analysis

Internal consistency was calculated by corrected item-total correlations, using the formula given by Nunnally, and by Cronbach’s alpha for each rater and overall.

Intrarater reliability was calculated using a two-way mixed effect ANOVA ICC model (3,1). The unit of analysis is an individual rating. Total variance is partitioned into between-subjects and error components. The formula used is as follows:

\[
\text{ICC} (3,1) = \frac{\text{BMS} - \text{EMS}}{\text{BMS} + (k - 1)\text{EMS}}
\]

where BMS is the between-subjects mean square from the analysis of variance, EMS is the error mean square, and \( k \) is the number of ratings for each subject. Shrout and Fleiss suggest that this model, in which the tested raters are considered the only raters of interest, is appropriate for testing intrarater reliability.

Interrater reliabilities were calculated using ICC model (2,1) which is based on a repeated measures ANOVA with raters as the independent variable. Variances are partitioned into effects due to differences between subjects, differences between raters, and error variance. ICC (2,1) is
used where all subjects (n) are measured by all raters (k) and the raters are considered representative of a larger population of similar raters. The unit of analysis is based on individual ratings. The F-ratio associated with the rater effect reflects the difference among raters, or the extent of agreement or disagreement among them. This effect is significant when the variance due to raters is large, indicating that the rater scores are different from each other. The formula used is as follows:

\[ \text{ICC}(2,1) = \frac{(\text{BMS} - \text{EMS})}{\text{BMS} + (k-1)\text{EMS} + k(\text{RMS} - \text{EMS})} \]

where BMS and EMS are between-subjects and error mean square respectively, k is the number of raters and n is the number of subjects.

Reliability coefficients and 95 percent confidence intervals were determined for the total score (expressed in percent) and across individual items. Statistical outputs were generated using SPSS version 10 software.

7.9 Pilot study results

7.9.1 Missing items

The maximum number of variables for the raw data was 576 (eight items per procedure times 12 ratings per procedure times six procedures). There were a total of 23 missing entries (4 percent) distributed among five of the items. There were no missing data for trocar insertion, camera/laparoscope control, and suction/irrigation. The most commonly missing item was hemostasis. These missing items were recorded as not observed and were interpreted to mean that the individual rater did not think that the videotaped segment for that item in terms of length and/or quality was adequate to render a judgement. Interestingly, one rater was responsible for 10 of the 23 missing entries. For the purpose of statistical calculation, the missing item score was replaced by the average score assigned to the remaining items by the particular rater for the particular procedure in which the item score was omitted.
7.9.2 Internal consistency

Table 2 provides the corrected item-total correlations across all raters and subjects for Time 1 and Time 2. Values ranged from 0.25 to 0.81.

<table>
<thead>
<tr>
<th>Item</th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar insertion</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>inspection of abdomen</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>scope control</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td>suction/irrigation</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>dissection</td>
<td>0.45</td>
<td>0.32</td>
</tr>
<tr>
<td>hemostasis</td>
<td>0.25</td>
<td>0.34</td>
</tr>
<tr>
<td>tissue removal</td>
<td>0.43</td>
<td>0.49</td>
</tr>
<tr>
<td>final inspection</td>
<td>0.35</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 3 displays Cronbach's alphas for each rater and across all raters for Time 1 and Time 2.

<table>
<thead>
<tr>
<th>rater</th>
<th>Cronbachs alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.91</td>
</tr>
<tr>
<td>B</td>
<td>0.90</td>
</tr>
<tr>
<td>C</td>
<td>-1.23</td>
</tr>
<tr>
<td>D</td>
<td>0.58</td>
</tr>
<tr>
<td>E</td>
<td>0.87</td>
</tr>
<tr>
<td>F</td>
<td>0.76</td>
</tr>
<tr>
<td>Overall raters Time 1</td>
<td>0.80</td>
</tr>
<tr>
<td>Overall raters Time 2</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Corrected Item-total correlations for individual raters are shown below in Table 4. Aside from rater C most items for most raters exceeded the minimal acceptable cut off point of 0.30. Exceptions were inspection (rater D), hemostasis (raters A,D), tissue removal (raters F,D) and Final inspection (rater B). Rater C exhibited negative correlations for seven of eight items. Rater C was clearly a perverse rater whose item intercorrelations were so low as to result in a negative Cronbach's alpha. The total variance for rater C's scores was 4.4, which was extremely low compared with the other raters and 79 percent (3.5) of this raters variance was accounted for by a single item, trocar insertion.

<table>
<thead>
<tr>
<th>Item</th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>rater D</th>
<th>rater E</th>
<th>rater F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar</td>
<td>.85</td>
<td>.97</td>
<td>-.03</td>
<td>.33</td>
<td>.90</td>
<td>.68</td>
</tr>
<tr>
<td>inspection</td>
<td>.61</td>
<td>.83</td>
<td>-.16</td>
<td>.24</td>
<td>.70</td>
<td>.51</td>
</tr>
<tr>
<td>scope cont</td>
<td>.96</td>
<td>.87</td>
<td>.34</td>
<td>.50</td>
<td>.98</td>
<td>.96</td>
</tr>
<tr>
<td>suct/irrig</td>
<td>.94</td>
<td>.82</td>
<td>-.09</td>
<td>.91</td>
<td>.82</td>
<td>.58</td>
</tr>
<tr>
<td>dissection</td>
<td>.83</td>
<td>.82</td>
<td>-.76</td>
<td>.44</td>
<td>.25</td>
<td>.73</td>
</tr>
<tr>
<td>hemostas</td>
<td>.01</td>
<td>.67</td>
<td>-.55</td>
<td>-.08</td>
<td>.89</td>
<td>.28</td>
</tr>
<tr>
<td>tiss remov</td>
<td>.84</td>
<td>.63</td>
<td>-.31</td>
<td>-.01</td>
<td>.64</td>
<td>-.22</td>
</tr>
<tr>
<td>final insp</td>
<td>.94</td>
<td>.18</td>
<td>-.22</td>
<td>.31</td>
<td>.62</td>
<td>.58</td>
</tr>
</tbody>
</table>

### 7.9.3 Rater Reliability

Table 5 displays the intrarater reliability and 95 percent confidence intervals for each rater across individual item and total LSI scores. For total scores, the values ranged from 0.35 to 0.93 with wide confidence intervals.
Table 5: Intrarater Reliability (ICC) and 95% confidence intervals for Individual Raters

<table>
<thead>
<tr>
<th>Item</th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>rater D</th>
<th>rater E</th>
<th>rater F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar</td>
<td>.61(-.19-.93)</td>
<td>.79(.17-.97)</td>
<td>.94(.70-.99)</td>
<td>-.47(-.89-.43)</td>
<td>1.00(1.0-1.0)</td>
<td>.88(.44-.98)</td>
</tr>
<tr>
<td>inspection</td>
<td>.87(.40-.98)</td>
<td>.53(-.30-.92)</td>
<td>.71(-.00-.95)</td>
<td>.39(-.45-.88)</td>
<td>.82(.26-.97)</td>
<td>1.00(1.0-.10)</td>
</tr>
<tr>
<td>scope cont</td>
<td>.71(.00-.95)</td>
<td>.71(-.00-.95)</td>
<td>.30(-.53-.86)</td>
<td>.69(-.05-.95)</td>
<td>.96(.80-.99)</td>
<td>.47(-.37-.90)</td>
</tr>
<tr>
<td>su/irrigat</td>
<td>.78(.14-.96)</td>
<td>.67(-.09-.94)</td>
<td>.46(-.37-.90)</td>
<td>.46(-.37-.90)</td>
<td>.88(.43-.98)</td>
<td>-.25(-.82-.61)</td>
</tr>
<tr>
<td>dissection</td>
<td>.77(.11-.96)</td>
<td>.55(-.28-.92)</td>
<td>.08(-.67-.78)</td>
<td>.19(-.61-.82)</td>
<td>.19(-.61-.82)</td>
<td>.41(-.43-.89)</td>
</tr>
<tr>
<td>hemostas</td>
<td>-.67(-.94-.17)</td>
<td>.63(-.16-.94)</td>
<td>.06(-.68-.78)</td>
<td>.69(-.05-.95)</td>
<td>.00(.00-.76)</td>
<td>.76(.10-.96)</td>
</tr>
<tr>
<td>tiss rem</td>
<td>.23(-.58-.84)</td>
<td>-.07(-.74-.72)</td>
<td>.03(-.70-.76)</td>
<td>.63(-.16-.94)</td>
<td>.62(-.17-.94)</td>
<td>-.11(-.76-.70)</td>
</tr>
<tr>
<td>final insp</td>
<td>.08(-.67-.78)</td>
<td>.09(-.67-.78)</td>
<td>.36(-.48-.87)</td>
<td>-.33(-.85-.55)</td>
<td>.08(-.67-.78)</td>
<td>.62(-.18-.93)</td>
</tr>
<tr>
<td>Overall</td>
<td>.73(.03-.96)</td>
<td>.82(.24-.97)</td>
<td>.35(-.49-.87)</td>
<td>.35(-.49-.87)</td>
<td>.93(.66-.99)</td>
<td>.75(.08-.96)</td>
</tr>
</tbody>
</table>
Table 6 displays the interrater reliability coefficients obtained for individual items and for overall scores for Time 1, Time 2 and Times 1 and 2 combined.

The interrater reliability for the total score (Time 1 and Time 2 combined) was ICC=0.47 (95% CI 0.16-0.86). The F-ratio associated with the rater effect was 4.6 (P=0.005), indicating a significant difference among raters.

<table>
<thead>
<tr>
<th>Item</th>
<th>Time 1</th>
<th>Time 2</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar insertion</td>
<td>.58 (.24-.90)</td>
<td>.54 (.22-.89)</td>
<td>.65 (.34-.93)</td>
</tr>
<tr>
<td>inspection</td>
<td>.55 (.23-.89)</td>
<td>.63 (.32-.92)</td>
<td>.64 (.33-.92)</td>
</tr>
<tr>
<td>scope control</td>
<td>.50 (.18-.87)</td>
<td>.44 (.14-.85)</td>
<td>.55 (.23-.89)</td>
</tr>
<tr>
<td>suction/irrigation</td>
<td>.34 (.08-.79)</td>
<td>.24 (.02-.73)</td>
<td>.32 (.07-.78)</td>
</tr>
<tr>
<td>dissection</td>
<td>.30 (.05-.77)</td>
<td>.23 (.00-.73)</td>
<td>.32 (.07-.78)</td>
</tr>
<tr>
<td>hemostasis</td>
<td>.08 (-.03-.50)</td>
<td>.01 (-.08-.38)</td>
<td>.07 (-.03-.47)</td>
</tr>
<tr>
<td>tissue removal</td>
<td>.00 (-.13-.46)</td>
<td>.22 (.02-.70)</td>
<td>.16 (.02-.65)</td>
</tr>
<tr>
<td>final inspection</td>
<td>.14 (-.00-.60)</td>
<td>.03 (.04-.38)</td>
<td>.11 (.00-.86)</td>
</tr>
<tr>
<td>Overall</td>
<td>.43 (.13-.84)</td>
<td>.38 (.10-.82)</td>
<td>.47 (.16-.86)</td>
</tr>
</tbody>
</table>

* F ratio, rater effect = 4.6 (p=.005)

7.9.4 Raters comments

There were no comments from the raters about difficulty in scoring the various items. However, as mentioned above, there were 23 missing entries (4 percent) distributed randomly among five of the items. It is easy to infer that, at least for a small number of videotaped segments,
there was a difference of opinion among raters as to whether a videotaped segment was adequately representational of the item to permit scoring.

Rater comments consisted of (1) questioning the justification for performing the operative procedure (2) explanations for scores assigned to particular items.

7.10 Pilot study interpretation of results

7.10.1 Internal consistency

While, overall, item-total correlations were satisfactory, the results obtained for individual raters were inconsistent and varied widely. The item correlating least with the total was hemostasis. The likeliest explanation for the low intercorrelations was ambiguity in defining each level of response as opposed to the item being too difficult, too easy, or unrelated to the attribute of interest. Among individual raters, the Cronbach's alpha levels ranged widely from poor to good. Failure to achieve a desired level of Cronbach's alpha may be due to (1) some items having close to zero correlation with the attribute (2) all items having low correlation or (3) a construct which is factorially complex so that the items are measuring several different attributes rather than one. However, in the case of the pilot study, it is likely that a fourth factor, insufficient clarification of the rating process and conversion criteria (see 8.2.3.1 and 8.2.3.2), played an important role in the widely differing values obtained.

7.10.2 Rater reliability

While intrarater reliability was in the moderate to good range, interrater reliability was in the poor to moderate range. Because interrater reliability contains all the sources of error contributing to intraobserver reliability, plus any differences which may arise between observers, the intrarater reliability is bound to be higher. An ICC of 0.47 for interrater reliability means that only 47 percent of variance in the scores resulted from "true" variance among subjects.
Two major reasons could explain the low ICC values. The first is that the raters did not agree. Because the ICC can be considered an average correlation across raters, it does not represent reliability of any individual rater. Thus, nonagreement may involve all raters, some raters, or only one rater. A second reason for a low (and sometimes negative) ICC may be a lack of variability among the subjects. This could occur when samples are homogeneous, or when raters are either very lenient or very strict in their scoring, or when the rating system falls within a restricted range (see 5.5). It is likely that both reasons contributed to the low interrater reliability in the pilot study. First, the rater effect was significant \( (p=0.005) \) and, second, the subjects selected came from a fairly restricted pool consisting of postgraduate trainees only, from PGY3 to PGY5, half of whom were at the PGY5 level.

The rater comments which questioned the justification for performing procedures were considered important in that the purpose of the LSI was not to evaluate judgement but rather, technical skill. In fact, the residents performing the procedures would not have been responsible for the operations selected. These choices would have been made by the surgeons in their offices prior to the patients ever coming to the operating room. Conceivably, if a rater did not feel the operation was justified, then this might have unfairly downgraded the scores by the rater for that particular procedure.

7.11 Summary

A pilot assessment of the acceptability and reliability of the LSI using videotapes was performed. The LSI was found to be easy to use, but internal consistency, based on Cronbach's alpha, ranged widely from poor to good for individual raters. Intrarater reliability was in the moderate to good range but interrater reliability was in the poor to moderate range. At this juncture, it was felt that the reliability of the LSI was insufficient to proceed to a study of construct validity.
CHAPTER 8: MEASURES TAKEN TO IMPROVE THE DESIGN, RATER PREPARATION AND CONTENTS OF THE LSI

Following a pilot study it is customary to make modifications that will improve the content. aggregation, instructions, criteria, or other arrangements of the index's construction and operational specifications. The next section deals with the modifications undertaken to improve the reliability of the LSI. This is followed by a description of the main reliability study and the results obtained.

8.1 Internal consistency

Cronbach’s alpha is dependent not only on the magnitude of the correlations among the items, but also the number of items in the index. Thus, an index can be made to look more "homogeneous" by rejecting items having low correlation with the total score and/or by increasing the number of items. In addition to adding new items until the desired level is achieved, item homogeneity might be resolved by rejecting items with low item-total correlation scores (≤0.20). The need by psychometricians to enlarge the number of items for statistical homogeneity may be at odds with the practicalities of ease of administration in a clinical setting.

For the main study, which follows, only one item, instrument knowledge, was deleted (see 8.2.3.2). This item had actually not been used in the pilot study for reasons previously stated. The other modifications made in the item list are described below in section 8.2.3.2. The modifications were not made to artificially lengthen the item list for the purpose of improving Cronbach’s alpha, but rather, for the sake of simplification and clarification.

8.2 Reproducibility

Reproducibility might improve following modifications to reduce the sources of measurement variability, namely (1) the variable being measured (subject) (2) the measuring
instrument (procedure) and (3) the individual taking the measurement (rater, observer, clinician)\textsuperscript{72,96}.

### 8.2.1 Subject variability

The subjects investigated should consist of individuals for whom the measurement is intended and the spectrum of performance should be sufficiently broad (heterogeneous) so as not to underestimate variability. The testing conditions should match the clinical setting. These principles were only partially met during the LSI pilot study. The subjects consisted only of residents within a restricted range of training (PGY3 to PGY5). Testing conditions were stable across all subjects. For the main reliability study in 1999 described below the sample size was greater, and to ensure that the procedures would reflect a broad range of surgical competency two maneuvers were used with respect to sample selection (1) half the procedures were performed in June, 1999 and half in July, 1999 following the July 1\textsuperscript{st} resident changeover. On July 1\textsuperscript{st} of each academic year, new residents are assigned to Women’s College to replace the outgoing residents. The July group would usually have less experience than the June group, particularly at the lower postgraduate levels. For example, PGY1’s beginning in July would have had no prior experience compared with PGY 1’s at the end of June who will have already had a full year of exposure to gynecologic surgery (2) faculty surgeons were included as subjects.

### 8.2.2 Procedure variability

The term \textit{procedure} used here refers to the process of videotaping and the videotaped product. During the pilot study and the main reliability study videotaping was performed in standard fashion as described in 7.5 and 9.8. Even though the number of missing items during the pilot study was low (four percent) efforts were made during the reliability study to further reduce this number by ensuring that the editing process contained adequate representational segments of each
videorecorded item, even if the videotaping for a particular procedure had to be extended.

8.2.3 Rater variability

There are generally two components to rater activities: the first step is the observational process. Guidelines for the performance of an observational process are called *process criteria*. The second step is the rater's conversion of the information into categories. The criteria used to make this conversion have been called *conversion criteria*.

Rater variability can be reduced by clarifying both the observational process and conversion criteria.

8.2.3.1 Process criteria

For the main reliability study rater instruction forms were revised with clarification of the objectives and the evaluation process (Appendix 7). An instructional video was produced with titles and still frames and samples of the various levels of skill performance. A pretrial meeting of the raters was convened to review the revised item list (see below 8.2.3.2), observe the instructional video and answer any questions. Because measurements vary less when performed under ideal circumstances, such as well rested observers, raters were instructed to not review more than two videotaped procedures at any one time.

8.2.3.2 Conversion criteria

Converting the raw information into categories (LSI ratings) was facilitated in the main study by partitioning the original eight item list into a 13 item list for simplification and clarification. The item *dissection* was broken down into three separate items: (1) *laser dissection* (2) *electrosurgical dissection* (3) *tissue and instrument handling*. This division seemed justified for two reasons. First, skill attainment with one modality does not imply the same level of skill attainment
with either of the others. For example, blunt dissection with scissors or the suction probe is substantially different from laser dissection. Second, it was easier to create more explicit descriptors by separating the types of dissection rather than combining them. The item tissue and instrument handling seemed to be a better term than blunt dissection because it incorporated the concept of appropriate choice and use of surgical instruments for dissection and expanded the scope of the item to deal not only with dissection, but the general care with which tissues are handled.

The item suction/irrigation used in the pilot study was separated into two items, suction and irrigation. Again, this enabled us to clarify the descriptors for what are two separate skills, suction and irrigation even though the same instrument, the suction/irrigation probe, is used to perform these two activities.

A new item, maintenance of visibility, was introduced into the item list because it was felt that visibility required manoeuvres which differed from laparoscope control, such as the ability to demonstrate a targeted area by proper retraction technique and appropriate camera adjustments.

In the main reliability study trocar placement replaced the term trocar insertion because the site(s) of trocar insertion seemed important in addition to the method of insertion. The item final inspection was replaced by procedure completion. The latter is more comprehensive because it includes not only inspection, but the method of clearing the abdomen of fluid, debris, gas and instruments.

One item, instrument knowledge, was dropped from the item list. It had not been used during the pilot study because instrument knowledge is not a videorecordable item. Moreover, to classify instrument knowledge as a technical skill, seemed inappropriate.
Thus, a revised item list was created for the reliability study consisting of 13 videorecordable items:

1. Veress needle/trocar placement
2. Initial inspection
3. Laparoscope control
4. Irrigation
5. Suction
6. Maintenance of visibility
7. Tissue and instrument handling
8. Electrosurgical dissection
9. Laser dissection
10. Knot tying/ligation
11. Hemostasis
12. Tissue removal
13. Procedure completion.

While the intent of the new item list was to facilitate the process of conversion criteria there was also the possibility that lengthening the item list would improve the internal consistency of the index. However, the author did not deliberately drop an item because of low item-total correlation nor was the item list extended for the sole purpose of improving the level of Cronbach’s alpha. Clarification and simplification were the primary reasons for the item changes in the revised list.

Other improvements in conversion criteria were undertaken. In the pilot study, the rater package contained a page of detailed descriptors separate from the LSI rating form in which the response categories for each item were labeled 1 to 5 with 1 being poor and 5 being excellent. In the main study, the number of response categories was maintained at five but the lowest level of performance was changed from poor to novice and the highest level from excellent to expert. Novice was described as an individual who continuously requires direction, performs the skill with hesitation and frequently requires correction. Expert was described as an individual who performs these skills smoothly in a controlled, progressive, self-directed fashion. It was felt that the use of
novice and expert would serve as better reference points for faculty raters than the terms poor and excellent.

In the main study, the rater package continued to have a separate sheet on which detailed descriptors of each item were provided (Appendix 8). However, unlike the pilot study, behaviourally anchored ratings (BARS) were placed subadjacent to the lowermost and uppermost rating categories of the scoring form to assist raters in the conversion process. Most research indicates that there is relatively little difference between scales with adjectives under each box versus end anchored scales. By using only end anchors, the tendency is to pull responses to the ends which are often ignored in unanchored scales. This promotes greater variability. Also, labeling every other box tends to result in more frequent endorsement of the labeled rather than the unlabeled boxes. Finally, the adjectival description for a mid category response is often difficult to create, poorly worded and ambiguous. (Appendix 8,9).

8.3 Summary

Various modifications were undertaken to improve the reliability of the LSI. These included improvement in the design format, rater preparation, item simplification and clarification and plans to obtain a sampling population which would be more representative of the ultimate target population of the LSI.
CHAPTER 9: MAIN LSI RELIABILITY STUDY ("MAIN STUDY")

9.1 Main study research questions

9.1.1 Primary research questions

1. To what extent do the items comprising the LSI correlate with each other and the total score achieved by summation of the individual item scores? That is, what is the internal consistency (homogeneity) of the LSI?

2. What is the interrater reliability of the LSI for individual items and total scores? That is, to what extent does the LSI exhibit similar results (reproducibility) when repeatedly applied to the same subject under constant conditions.

9.1.2 Secondary questions

3. Following the measures taken to improve the design, rater preparation and contents of the LSI, was there an observed improvement in both internal consistency and interrater reliability in the main study?

4. Is the level of internal consistency sufficient to describe the LSI as a unidimensional index?

5. Is the level of interrater reliability sufficiently high to proceed to a study of construct validity of the LSI as a means of objectively assessing laparoscopic skills among gynecologists?

9.2 Main study summary of methods

The field work for the main study was conducted in June and July, 1999 following measures implemented to improve the reliability of the LSI. Videotaped records of 20 gynecologic laparoscopic procedures were edited to approximately 30 minutes in length or what was necessary
to capture representational segments of the videorecordable LSI items. The 20 procedures were compiled onto five videotapes and four raters reviewed each compilation.

9.3 **Main study subject (procedure) selection**

All procedures were performed by postgraduate residents and faculty in gynecology at Women's College Hospital. All resident procedures were supervised by faculty members.

9.3.1 **Main study procedure eligibility criteria**

The laparoscopic procedures used for the pilot study met the following eligibility criteria:

1. Surgery was booked for no less than 60 minutes.
2. The procedure was booked in a designated gynecologic laparoscopic operating room.
3. The procedures were performed on an elective basis.
4. All operations qualified as level II laparoscopic procedures in accordance with the guidelines for training in laparoscopic procedures prepared by the Society of Obstetricians and Gynaecologists of Canada.\(^{32}\)

9.3.2 **Main study procedure exclusion criteria**

The following were excluded from the study:

1. Nonelective (emergency procedures)
2. Level I procedures such as diagnostic laparoscopy and laparoscopic sterilization, and level III procedures such as the Burch sling for urinary stress incontinence or radical hysterectomy and lymphadenectomy.
3. Procedures where the supervising faculty rater was one of the pilot study raters.

9.4 **Main study rater selection**

The four raters were selected from the initial cadre of 13 content experts. Two raters were
from Women’s College, one from The Toronto Hospital and one from The Mississauga-Queensway Hospital - Queensway Site.

9.5 **Main study sample size calculation**

Confidence interval formulae for ICC’s are complex and difficult to use for sample size estimation. An alternative approach for sample size determination is provided by Donner and Eliasziw (1987)\(^{106}\). They based the calculation on the number of subjects required to determine if the coefficient is significantly different from some arbitrary value. They developed exact power contours to guide the planning of reliability studies where the parameter of interest is the coefficient of intraclass correlation (\(\rho\)), derived from an ANOVA model. The contours display the required number of subjects (\(k\)) and the number of repeated measurements (\(n\)) that provide 80 percent power for testing the hypothesis \(H_0: \rho \leq \rho_0\) versus \(H_1: \rho > \rho_0\) at the 5 percent level of significance for selected values of \(\rho\). In a more recent paper, Walter, Eliasziw and Donner (1998) developed a simple approximation that allowed the calculation of required sample size for the number of subjects (\(k\)) when the number of replicates (\(n\)) is fixed \(^{107}\). For the main study, the interrater reliability coefficient obtained in the pilot LSI study (0.47) was rounded to 0.50. Using this value as \(\rho_0\), and four replicates (raters), the sample size required to demonstrate \(\rho > \rho_0\) based on alpha=0.05 and beta=0.20 would be 12 for a desired \(\rho\) of 0.80. Thus, the difference between moderate and very good would be demonstrated with fewer than 20 subjects and four raters.

For \(H_0: \rho_0 = 0.50\) and \(H_1: \rho = 0.80\), with alpha=0.05 and beta=0.20 and three raters, the required sample size would be 14. Portney and Watkins state that values above 0.75 for reliability coefficients are “good to excellent”\(^{73}\). Reliability coefficients in this range would be sufficient to proceed to a validity study. Thus, the sample size required would be 12 subjects when 4 raters are
used and 14 subjects for 3 raters. A sample size of 20 was chosen for this study to ensure a broad spectrum of types of operations and levels of surgical competency and to compensate for dropouts and lost data.

9.6 Main study feasibility

Approximately 15 gynecologic laparoscopic procedures are performed at Women's College Hospital each week. Videolaparoscopy equipment is kept in the laparoscopy suites and is available for all operative procedures. During the pilot study, six videotaped procedures were obtained within a two week time frame. It was anticipated that videocapture of 20 procedures could be accomplished within a six week period.

9.7 Main study ethics approval

The study was submitted to the Research Ethics Board of Sunnybrook and Women's College Health Science Centre, Women's College Campus. Information sheets and consent forms were developed for the patients, participating faculty surgeons and postgraduate trainees. (Appendix 10). The procedure videotapings were for the determination of LSI rating scores only and were not to be used in any manner to guide treatment or alter management of the involved patients. The surgeons (faculty and residents) and patients participating in this study remained anonymous with no videodisplay of surgeon or patient names or procedure dates. All anonymous tapes were kept in the department teaching files following completion of the study. The LSI scores obtained were not to be used in the trainees end-of-rotation ITER reports. The study was approved by the Research Ethics Committee on June 9th, 1999 (Appendix 11).
9.8 **Main study videorecordings and videotape compilations**

Prior to commencing the field work for the main reliability study the investigators circulated a letter to the gynecology staff at Women’s College Hospital apprising them of the planned study and requesting their assistance with recruiting subjects for videotaping procedures (Appendix 12).

Twenty procedures fulfilling the eligibility criteria were videotaped on a first come-first serve basis following the approval of the Research Ethics Committee. In 10 cases the subjects were postgraduate trainees ranging from PGY1 to post-certification fellow, and in the remaining 10 the subjects were members of the gynecology faculty at Women’s College Hospital.

All procedures were electively performed during day time operating hours in one of two dedicated gynecologic laparoscopy suites. The operating room nursing staff in these suites were all trained for gynecologic laparoscopy and were accredited in laser instrumentation. Operating room standards, surgical monitoring equipment (including videocapture) were standardized. The surgery was performed in the usual fashion for the operating surgeon with no modifications for the purpose of videorecording. The images were transmitted to a monitor as usual and there were no audio components. Videorecordings were made with high quality super VHS tapes (TDKXP superpro/ST120, Fuji S-VHS/ST160) using S-VHS recorders (Sony). The videotaped procedures were edited by one of the investigators (RP). Representative segments of videorecordable LSI items were included. The video product was shorter than the actual operating time by eliminating "downtime" (periods where operative activity is suspended, for example, during attachment of the laser or replacement of irrigation fluids) and duplications (for example, if bilateral salpingoophorectomies were performed, the videoediting selected one side only). The videotaped procedures were compiled with titles clearly defining when one operation began and another ended.
The operative procedures were arranged in chronologic order, however, the order seen on the final compilation was determined by computerized random number generation. Only intraabdominal events were visible on the videotapes.

9.9 Main study rating procedure

In preparation for scoring, a revised multicoloured rater package was prepared. It consisted of four pages (1) a revised rater instruction form (2) two pages of item descriptors (3) a single page LSI rater form (Appendices 7,8,9). On September 10th, 1999 after the videotaped procedures had been compiled, a meeting was convened of the investigators and raters. At that meeting the rater packages were distributed and the contents were reviewed with the raters. An instructional video which had been produced to demonstrate the range of skill levels from novice to expert was shown. All issues and questions were dealt with. The raters were directed to review their respective compilations within a two week period, but to avoid a fatigue factor, a maximum of two procedures were to be reviewed per day. The raters were advised to complete the LSI rater form for each operative procedure as soon as the procedure had been reviewed in its entirety.

9.10 Main study data collection and entry

The videotapes and completed LSI forms were collected at the end of the two week period from each rater and the raw data were entered into a Microsoft Excel 97 worksheet by one of the investigators (RP) and doubly checked for correctness by the other investigator (JS).

9.11 Main study outcomes

9.11.1 Internal consistency

Internal consistency was measured by Cronbach’s coefficient alpha for individual raters and for all raters combined. Corrected item-total correlations were determined for each rater and for all
raters combined using the formula given by Nunnally and the Pearson Product Moment Correlation.

**9.11.2 Rater reliability**

Interrater reliability was calculated using the repeated measures two-way random effects ANOVA to generate an ICC. The statistical equation used was ICC (2,1), the model used where all subjects are measured by all raters and the raters are considered representative of a larger population of similar raters. ICC (2,1) is used when single ratings from individual raters are compared. Interrater reliabilities and 95 percent confidence intervals were calculated for each item and for the final scores (percent) across all procedures (subjects).

**9.12 Main study results**

Although the LSI used in the main study consisted of 13 items, two items were dropped prior to conducting the statistical analyses. One of these, knot tying/ligation was not observed during any procedure by any rater. This was identical to the pilot study, reflecting the fact that knot tying/ligation in level II gynecologic laparoscopy procedures is very infrequently performed. Thus, it has very little role to play in an LSI designed for level II gynecologic procedures. In **CHAPTER 10: DISCUSSION**, the role of knot tying/ligation as an item in potential future versions of the LSI is discussed.

Electrosurgical dissection was scored only 20 times. Rater A was responsible for 12/20 entries and the other three raters combined scored electrosurgical dissection only eight times. Thus, it was felt that rater A was including under electrosurgical dissection events that did not qualify, such as bipolar coagulation of a pedicle prior to dividing it or electrosurgical techniques for hemostasis. Rater A, the community rater, differed from the three university raters, B, C, and D in other respects and this is discussed below. Therefore, because electrosurgical dissection was very
infrequent and because the numbers for raters B, C and D were very small, electrosurgical dissection was also dropped from the final analysis. At Women's College Hospital, the laser modality was used almost exclusively. Whether or not to retain electrosurgical dissection in future versions of the LSI is addressed in CHAPTER 10: DISCUSSION.

After removing electrosurgical dissection and knot tying/ligation, the LSI contained a maximum of 11 items for statistical analysis. For some procedures, there were fewer items when a particular skill, such as tissue removal or hemostasis, was not required.

9.12.1 Management of various data issues prior to statistical analysis

9.12.1.1 Observational conflicts

1. If three of four raters reported that an item was not observed, then the remaining rater’s entry was judged to be a false positive and the result was converted to not observed.

2. If three of four raters scored an item and the fourth rater marked it as not observed then the latter’s entry was considered to be a false negative. The videotapes were redistributed to the raters who had marked the item as not observed. They were asked to assign a score if, on reviewing the tapes, it was felt to be reasonable to do so.

3. When the raters were equally split as to whether an item was present or absent, the situation was managed as follows: the investigators reviewed the videotapes and made a judgement as to whether the item was present or absent. If the judgement was that the item was absent, the scores assigned by two of the raters were considered to be false positives and were converted to not observed. If the item was judged to be present, then the not observed entries were considered to be false negatives and the videotapes were returned to the raters who had marked the item as not observed. They were asked to assign a score if, on
reviewing the tapes, they felt they could reasonably do so.

9.12.1.2 False positives

There were nine false positive scores, eight for hemostasis and one for tissue removal. Seven of the nine false positive entries belonged to rater A. The false positives were seen in seven cases.

9.12.1.3 False negatives

There were twelve false negatives, eight for tissue removal, one for hemostasis, one for laser dissection, one for suction, and one for procedure completion. Rater A was responsible for six of the twelve false negatives.

In all instances of false negative entries, the raters were able to assign a score for the missing item on reviewing the appropriate videotape. After correcting for false positives and false negatives, the distribution of the number of items per case was as follows:

11 item cases-9
10 item cases-5
9 item cases-6

The final number of scores available for analysis was 812 (203 items times 4 raters). Had 11 items been present on each videotaped procedure, the maximum number of item scores would have been 880. The reduction occurred as follows: 1) there was full agreement by all four raters that tissue removal did not take place in six cases (24 scores), hemostasis in two cases (8 scores) and laser in two cases (8 scores) 2) a further reduction of 28 scores followed removal of the seven items in which false positive scores occurred. Overall, there were scoring "errors" (false positives and false negatives) in 21 of 312 (2.6 percent) of scores.

9.12.2 Analysis options
Analyses were done in different ways. Internal consistency and rater reliability were initially calculated separately for 11, 10 and 9 item cases (designated as the Raw Data Set). Subsequently, not observed items in each case were replaced by the average score obtained on the remaining items for that particular case and rater. Thus, all procedures became 11 item cases (880 data points) and the analyses were repeated (Full Data Set).

Figure 1, Appendix 13 graphically displays the total scores (percent) assigned by each rater to each subject. The trend line for rater A (the community rater) is higher than the trend lines for the other three raters (university raters) which are all closely cropped together. Thus, rater A appears to be a more lenient scorer than raters B, C, and D. The mean scores, standard deviations and range of scores for each rater, shown in Table 7 below, further demonstrate that rater A is a more lenient rater and uses a rating system that falls within a more restricted range. Because of the outlier position of rater A compared with raters B, C, and D, all analyses were performed in two different manners (1) across all four raters (2) across raters B, C, and D only.

<table>
<thead>
<tr>
<th>Individual Raters</th>
<th>mean score (%)</th>
<th>std dev</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>84.75</td>
<td>9.16</td>
<td>67 to 96</td>
</tr>
<tr>
<td>B</td>
<td>68.00</td>
<td>21.37</td>
<td>33 to 100</td>
</tr>
<tr>
<td>C</td>
<td>61.35</td>
<td>18.98</td>
<td>35 to 92</td>
</tr>
<tr>
<td>D</td>
<td>66.4</td>
<td>18.66</td>
<td>31 to 87</td>
</tr>
</tbody>
</table>

While raters, B, C, and D showed closely cropped trend lines and scores for the last 18 of 20 cases, it is apparent from Figure 1 that the scores obtained in cases 1 and 2 for rater D did not approximate the scores obtained by raters B and C which were close to one another. Theoretically,
this difference may be related to a "warm-up effect" \(^{73,96}\), reflecting a lack of practice during the initial phase of the trial. Therefore, it was decided also to conduct analyses first on all 20 subjects and then on the last 18 of 20 subjects to see what level of reliability was achieved following a possible initial warm-up effect.

9.12.3 Internal consistency

Table 8 shows the mean scores and standard deviations for individual items and 1) all raters combined 2) raters B, C and D excluding rater A. The lowest mean score was seen with tissue and instrument handling, however, none of the items stood out as being particularly easy or difficult.

<table>
<thead>
<tr>
<th>Item</th>
<th>mean all raters</th>
<th>std dev all</th>
<th>mean B,C,D</th>
<th>std dev B,C,D</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar plac</td>
<td>3.50</td>
<td>1.22</td>
<td>3.24</td>
<td>1.21</td>
</tr>
<tr>
<td>init insp</td>
<td>3.70</td>
<td>1.23</td>
<td>3.43</td>
<td>1.33</td>
</tr>
<tr>
<td>scope cont</td>
<td>3.48</td>
<td>1.10</td>
<td>3.20</td>
<td>1.11</td>
</tr>
<tr>
<td>irrigation</td>
<td>3.74</td>
<td>1.08</td>
<td>3.50</td>
<td>1.11</td>
</tr>
<tr>
<td>suction</td>
<td>3.41</td>
<td>1.22</td>
<td>3.06</td>
<td>1.27</td>
</tr>
<tr>
<td>maint of visib</td>
<td>3.40</td>
<td>1.18</td>
<td>3.15</td>
<td>1.25</td>
</tr>
<tr>
<td>tiss &amp; inst hand</td>
<td>3.28</td>
<td>1.30</td>
<td>2.94</td>
<td>1.31</td>
</tr>
<tr>
<td>laser dissection</td>
<td>3.42</td>
<td>1.24</td>
<td>3.12</td>
<td>1.29</td>
</tr>
<tr>
<td>hemostasis</td>
<td>3.45</td>
<td>1.16</td>
<td>3.21</td>
<td>1.17</td>
</tr>
<tr>
<td>tissue removal</td>
<td>3.40</td>
<td>1.06</td>
<td>3.10</td>
<td>1.00</td>
</tr>
<tr>
<td>proc comp</td>
<td>3.78</td>
<td>1.09</td>
<td>3.59</td>
<td>1.17</td>
</tr>
</tbody>
</table>

Item-total correlations and Cronbach's alphas for individual raters and for all raters combined are displayed in Table 9.
Table 9: Corrected Item-Total Correlations and Cronbach's Alphas for Individual Raters and for All Raters Combined (Main Study)

<table>
<thead>
<tr>
<th>Item</th>
<th>rater A</th>
<th>rater B</th>
<th>rater C</th>
<th>rater D</th>
<th>all raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar pl</td>
<td>.46</td>
<td>.60</td>
<td>.53</td>
<td>.74</td>
<td>.63</td>
</tr>
<tr>
<td>init insp</td>
<td>.61</td>
<td>.86</td>
<td>.76</td>
<td>.68</td>
<td>.76</td>
</tr>
<tr>
<td>scope cont</td>
<td>.76</td>
<td>.88</td>
<td>.90</td>
<td>.94</td>
<td>.86</td>
</tr>
<tr>
<td>irrigation</td>
<td>.56</td>
<td>.92</td>
<td>.86</td>
<td>.91</td>
<td>.86</td>
</tr>
<tr>
<td>suction</td>
<td>.56</td>
<td>.89</td>
<td>.71</td>
<td>.76</td>
<td>.79</td>
</tr>
<tr>
<td>maint of visibility</td>
<td>.69</td>
<td>.90</td>
<td>.90</td>
<td>.87</td>
<td>.84</td>
</tr>
<tr>
<td>tiss &amp; inst handling</td>
<td>.73</td>
<td>.85</td>
<td>.79</td>
<td>.81</td>
<td>.84</td>
</tr>
<tr>
<td>laser dissec</td>
<td>.48</td>
<td>.88</td>
<td>.62</td>
<td>.89</td>
<td>.81</td>
</tr>
<tr>
<td>hemostasis</td>
<td>.86</td>
<td>.90</td>
<td>.88</td>
<td>.90</td>
<td>.91</td>
</tr>
<tr>
<td>tissue removal</td>
<td>.60</td>
<td>.78</td>
<td>.43</td>
<td>.67</td>
<td>.70</td>
</tr>
<tr>
<td>proc comp</td>
<td>.31</td>
<td>.88</td>
<td>.79</td>
<td>.79</td>
<td>.73</td>
</tr>
<tr>
<td>C alpha (raw data)</td>
<td>.86</td>
<td>.97</td>
<td>.93</td>
<td>.96</td>
<td>.95</td>
</tr>
<tr>
<td>C alpha (full data)</td>
<td>.88</td>
<td>.97</td>
<td>.94</td>
<td>.96</td>
<td>.95</td>
</tr>
</tbody>
</table>

Item-total correlations for all items and all raters exceeded the minimal cutoff point of $r = 0.30$ and, compared with the pilot study, item-total correlations improved across all items. Cronbach's alpha also improved for individual raters and for all raters combined, exceeding the level of 0.90 considered to be the minimal satisfactory level of homogeneity for decisions pertaining to individual subjects.

9.12.4 Rater reliability

The ICC's and 95 percent confidence intervals for interrater reliability, using four different combinations of raters and subjects, are depicted in Table 10. When rater A is excluded the
reliability coefficient is in the good range and when only the last 18 subjects are analyzed across raters B, C, and D, the reliability coefficient is in the good to excellent range. Even with 18 subjects and three raters, the power of the study necessary to demonstrate $H_1: \rho > \rho_0$ was not diminished. The lowest value obtained for the overall ICC occurred when four raters across 20 subjects were analyzed. The result was 0.46 (95% CI 0.19 to 0.71). The reliability coefficient rose to 0.69 (95% CI 0.47 to 0.85) if rater A was excluded from the analysis. If rater A was excluded and only the last 18 subjects were considered (that is, after a warm up period) the ICC was 0.77 (95% CI 0.56 to 0.90).

The F-ratio for three raters was 2.0 ($p > 0.05$). With all four raters the F-ratio was 16.8 ($p < 0.001$), reflecting the systematic bias associated with rater A.

<table>
<thead>
<tr>
<th>Table 10: Interrater Reliability (ICC) and 95% Confidence Intervals for Different Combinations of Raters and Subjects (Main Study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combinations of Raters and Subjects (Main Study)</td>
</tr>
<tr>
<td>raters</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>ABCD</td>
</tr>
<tr>
<td>ABCD</td>
</tr>
<tr>
<td>BCD</td>
</tr>
<tr>
<td>BCD</td>
</tr>
</tbody>
</table>

* Subjects 3 to 20

Table 11 displays ICC levels and 95 percent confidence intervals obtained with combinations of four and three raters and 20 and 18 subjects, for individual items.
### Table 11: Individual Item ICC’s and 95 % Confidence Intervals for Different Combinations of Raters and Subjects (Main Study)

<table>
<thead>
<tr>
<th>Item</th>
<th>raters/subjects 4*/20</th>
<th>raters/subjects 4/18</th>
<th>raters/subjects 3*/20</th>
<th>raters/subjects 3/18</th>
</tr>
</thead>
<tbody>
<tr>
<td>trocar pl</td>
<td>.31 (.07-.58)</td>
<td>.32 (.07-.61)</td>
<td>.34 (.05-.63)</td>
<td>.37 (.04-.67)</td>
</tr>
<tr>
<td>init insp</td>
<td>.49 (.27-.72)</td>
<td>.52 (.28-.75)</td>
<td>.72 (.52-.87)</td>
<td>.77 (.58-.90)</td>
</tr>
<tr>
<td>scope control</td>
<td>.43 (.18-.68)</td>
<td>.49 (.21-.73)</td>
<td>.55 (.28-.57)</td>
<td>.62 (.36-.82)</td>
</tr>
<tr>
<td>irrigation</td>
<td>.38 (.16-.63)</td>
<td>.44 (.20-.69)</td>
<td>.65 (.41-.83)</td>
<td>.74 (.52-.82)</td>
</tr>
<tr>
<td>suction</td>
<td>.39 (.15-.64)</td>
<td>.38 (.12-.65)</td>
<td>.52 (.27-.76)</td>
<td>.51 (.19-.76)</td>
</tr>
<tr>
<td>tiss &amp; inst hand</td>
<td>.32 (.11-.57)</td>
<td>.37 (.13-.64)</td>
<td>.57 (.32-.78)</td>
<td>.67 (.44-.85)</td>
</tr>
<tr>
<td>maint of visib</td>
<td>.43 (.21-.67)</td>
<td>.47 (.22-.72)</td>
<td>.61 (.36-.81)</td>
<td>.69 (.41-.86)</td>
</tr>
<tr>
<td>laser dissection</td>
<td>.31 (.10-.58)</td>
<td>.35 (.12-.62)</td>
<td>.43 (.17-.69)</td>
<td>.50 (.22-.75)</td>
</tr>
<tr>
<td>hemostasis</td>
<td>.48 (.20-.72)</td>
<td>.51 (.21-.75)</td>
<td>.74 (.55-.88)</td>
<td>.78 (.58-.90)</td>
</tr>
<tr>
<td>tissue removal</td>
<td>.23 (-.05-.49)</td>
<td>.31 (.09-.58)</td>
<td>.59 (.34-.79)</td>
<td>.38 (.10-.66)</td>
</tr>
<tr>
<td>proc comp</td>
<td>.43 (.21-.67)</td>
<td>.45 (.22-.70)</td>
<td>.25 (-.02-.55)</td>
<td>.62 (.36-.82)</td>
</tr>
</tbody>
</table>

*raters A,B,C,D

Individual item ICC’s varied widely, generally in the moderate to good range with better results when only 3 raters where considered ,especially if only the last 18 subjects included in the analysis. Thus, in general, the rater reliabilities of the main study were greater than the results achieved in the pilot study. This improvement may have resulted from the various measures implemented to reduce sources of variability. Rater A, the community rater, was a more lenient rater and worked within a restricted range compared with raters B, C, and D. The latter three raters, the university raters, exhibited a high level of agreement across 20 subjects and, if one accepts a warm up effect, an even greater level of agreement across the last 18 subjects.
9.13 Summary

A main reliability study was undertaken in which four raters each assessed videotaped operations performed by 20 subjects. The subjects were a heterogeneous population comprised of various levels of postgraduate trainees and gynecologic faculty. The results of the main study indicated that the LSI now had a high level of internal consistency. After excluding a single rater who demonstrated systematic bias in the form of more lenient scores within a restricted range, interrater reliability was found to be in the good to very good range.
CHAPTER 10: DISCUSSION

10.1 Introduction

The goal of this thesis was to construct, and test the reliability, of the LSI, a new, objective measure of level II gynecologic laparoscopy skills in the human setting. The construction phase included (1) item selection (2) selecting a method to scale responses (3) design (4) choosing a scoring method. The components of reliability, internal consistency and reproducibility, were then evaluated. A pilot study was first conducted in which six raters each evaluated six subjects using edited videotapes of operative procedures performed by gynecology residents. Following the pilot study various modifications were made in an effort to improve reliability. A main reliability study was then undertaken in which four raters assessed 20 videotaped laparoscopic procedures performed by gynecology trainees, residents and faculty.

The results of the main study indicate that the LSI has a high level of internal consistency. In particular, Cronbach's alpha for individual raters and all raters combined exceeded the level of 0.90 considered to be the minimal satisfactory level of homogeneity for decisions pertaining to individual subjects. Thus, the LSI appears to be a multiitem, unidimensional index in which the item variables represent true components of the overall attribute. The high alpha levels achieved may reflect some item redundancy and this opens the possibility for the inclusion of fewer items in the index. However, the LSI is already a brief index and easy to administer. By reducing the number of items there is a risk that the domain of content will not be fully covered. Also, it is likely that there will always be some element of redundancy in evaluating laparoscopic or surgical skills. The competence with which one skill is performed is likely to affect the level of skill attained on other items. For example, laser dissection is likely to be performed best when the maintenance of visibility is optimal. The latter, in turn, may depend on adequate retraction of intervening structures.
by the use of proper trocar placement.

In the main study interrater reliability improved when rater A, whose scores were systematically higher than the others, was excluded from the analysis. If only the last 18 of 20 subjects were analyzed (assuming that the first two subjects constituted a rater "warm up" period) the reliability coefficient rose even higher. The number of subjects and raters required to demonstrate $H_1: \rho > \rho_0$ were 14 and 3 respectively, if $\rho_0$ was 0.50 and the desired level of $\rho$ was 0.80. The reliability coefficients obtained in the main study for three raters and 20 subjects is considered good, and for three raters and 18 subjects, it is considered to be very good to excellent. It should be noted that ICC model used (2,1), which is based on individual rather than mean ratings, is usually associated with the lowest reliability coefficient. For example, had each rater's score represented a mean of 2 or more ratings, the reliability coefficient for three raters and 18 subjects would have been 0.91. If the analysis had been performed using ICC model (3,k) in which a mean rating is obtained for fixed raters, the reliability coefficient across raters B, C, and D would have been 0.92. However the LSI is intended to be used by different raters in the future. Thus, it was appropriate to use ICC model 2 rather than model 3 for the analysis.

In summary, the results of the main study provided evidence that the LSI has a very high level of internal consistency, and excluding the systematic bias of rater A, good to very good interrater reliability. These results represent a clear improvement over those obtained in the pilot study. In the author's opinion the improvements can be ascribed to the measures described in \textbf{CHAPTER 8: MEASURES TAKEN TO IMPROVE THE DESIGN, RATER PREPARATION AND CONTENTS OF THE LSI}, in particular the pretrial rater instructional meeting and the redesign of the scoring form to include adjectival end anchors. These findings justify proceeding to a study to evaluate construct validity of the LSI (see below, \textbf{Future Research}).
10.2 Prevention and management of bias in the LSI reliability studies

In designing both the pilot and main studies, efforts were made at the outset to reduce or eliminate various potential biases.

Because raters can be influenced by the memory of the first score, there is a possibility for recall bias when the same rater is involved in both the test and retest. This was a risk in the pilot study assessment of intrarater reliability. Ordinarily, intervals between testing should be far enough apart to avoid fatigue, learning, or memory effects, but close enough to avoid genuine changes in the measured variable. A retest interval of 2-14 days is usual. In the pilot study there was no reason to be concerned about a change in the measured variable over time since the observed procedures on the two occasions were identical videorecordings. Thus, to err on the side of safety, a one month interval was selected in the pilot study to override the effects of recall. Further, to reduce the likelihood of recall bias, the procedure orders in time 1 and time 2 videotaped compilations were different.

Because of the subjective element intrinsic to a global assessment rater bias might be a concern for two reasons (1) a rater could assign a score based on his/her prior impression of the surgeon being assessed (2) the rater might be influenced by his/her judgement of the need for, or type of procedure, being performed. The first concern was not an issue because videotapes show only intraabdominal events making it impossible for the raters to identify the subjects. Also, the names of surgeons and patients were omitted from the videotapes. The second concern was more relevant. Rater comments on the pilot study scoring forms which queried the justification for, or choice of, a particular procedure, raised the possibility that raters might have been influenced to downgrade a subject's score in certain instances. In the main study, the rater preparation and instruction form emphasized the need to separate the quality of technical skills observed from the
other aspects of surgical competence, such as clinical decision making and the choice of procedure. Scales which are scored on a continuum, such as visual analog and Likert scales, are prone to other types of biases. These include end-aversion bias, positive skew, and the halo effect. To overcome end-aversion bias, which refers to the reluctance of some people to use the extreme categories of a scale, adjectival anchors subadjacent to the lowest and highest score for each item were introduced in the main study version of the LSI. End-anchored scales tend to pull responses towards the ends, producing greater variability and overcoming a central tendency bias.

There is also a tendency for assessments of clinical skills to be skewed towards superlative ratings. For example, Linn (1979) found that the mean score on a five point scale was 4.11 rather than 3.00, and the scores ranged between 3.30 and 4.56. Thus, the lower half of the scale was never used. The effect of this skew is to produce a ceiling effect with most of the marks clustered in a few boxes at one extreme. This was not a factor in the pilot study where the mean score for items was 3.25 and there was no evidence of a clustering of the scores in either half of the scale. Nor was it a factor in the main study in which the mean score for items was 3.3. The absence of skewing was likely due to several effects (1) blinding of the rater to the surgeon’s identity (2) an awareness by the raters that the videotapes were not to be used for purposes outside of the research study, such as the ITER (3) the use of BARS in the main study which would serve as concrete reference points for raters (4) adjectival end anchors in the main study and (5) instructions to the raters during the main study which emphasized the need to evaluate each item on its own merits rather than applying a “halo effect”, wherein one high scoring item would influence the other item scores.
In the main study there was a **systematic bias** associated with rater A who was clearly more lenient, and worked within a restricted range, compared with raters B, C, and D. For this reason, the analyses were performed with and without rater A. The strategies for dealing with extreme observers include eliminating the observer from further involvement or adjusting to the error by adding or subtracting the appropriate constant $^{63,73,110}$. Even more to the point would be the prevention of systematic error by rater training programs which clearly define the goals of assessment and provide explicit instructions for the rating process.

### 10.3 Limitations of the research

There are two potential limitations of this research which merit discussion.

The first of these relates to the videotaped product. In 4.4.5 the pros and cons of videotape administration of the LSI were discussed. The disadvantages of videotaping relevant to the research protocol were (1) inability of the rater to tell which part of the procedure was done by the surgeon being assessed (2) the videoediting process in which item(s) might not be adequately captured to render an assessment and (3) exclusion of nonvideorecordable items.

Because it is impossible for a blinded rater to separate the surgeon from the assistant, it would actually be more appropriate to state that a procedure, rather than a specific surgeon, was being assessed. This should not have altered the interrater reliability in the experiment. However, in future applications of the LSI, when rating is performed using videotapes, raters must be assured that the portion of the procedure being assessed actually represents the target subject on whom an assessment is desired. If the rater was also the supervising surgeon at the operation there would be no problem in differentiating the surgeon from resident. On the other hand, if a third party is reviewing the videotape, then it would be important to delineate, either by notation or the
videoediting process, exactly who is doing what.

Regarding the videoediting process, there were a small number of instances in both studies where the raters disagreed about the presence or absence of an item (see 9.12.1.2 and 9.12.1.3). In the pilot study, missing items were replaced by an average of the remaining items for the particular procedure and rater. In the main study conflicts were resolved by eliminating false positives (at the investigators’ discretion) and, in the case of false negatives, by asking raters to assign a rating, if at all possible, after a review of the videotape. In each instance the rater assigned a score but if this action merely reflected a willingness to be agreeable, the score assigned may not have represented his/her true opinion about the presence or absence of the item. In any event, these instances were very few and did not alter the statistical results. In part, the limitations of the videoediting process related to the fact that raters were being asked to devote a considerable amount of their personal time by participating in the study, and so an attempt was made to balance the length of the videoedited product with the various raters’ time constraints. In future applications, where reviewing LSI videotapes will not have to be performed in large quantities over a very limited time period, it will be important to ensure that the videoediting process does not compromise the length of representational segments for the various items.

Another limitation of videotaping product pertains to the fact that certain skill events take place outside the abdomen and are, therefore, not videorecordable. These include Veress needle insertion and extracorporeal knot tying. Veress needle insertions are a primary cause of abdominal injury during laparoscopy and it is unknown whether trocar placement can serve as a surrogate item for the Veress needle insertion. This will have to be tested in a future study comparing videotape and in vivo assessments of the LSI. Extracorporeal knot ligation is a skill which would ordinarily
be seen in level III laparoscopic procedures. The delivery of an extracorporeal knot to the targeted site is videorecordable but the actual knot tying and insertion of the knot into the introducer is not. Therefore, if a modified version of the LSI is used in the future for level III procedures, it will be necessary to determine the extent to which extracorporeal maneuvers affect the final score and whether or not videotape administration of the LSI is as valid for these procedures as in vivo assessments.

A second possible limitation of this research relates to the theory of generalizability. According to generalizability theory, reliability should be interpreted within a multidimensional context, that is, in relation to a set of specific testing conditions. Each condition that defines this context is called a facet. A particular combination of facets characterizes the universe to which reliability can be generalized. Every individual score can be thought of as a sample from a universe of possible scores that might have been obtained under the same testing conditions. To make generalizations based on the reliability studies, it is necessary to assume that the sample of subjects, the conditions for collecting data and the range of skills are similar to the eventual target population. Also, the raters chosen should be representative of the pool of raters who will ultimately use the index. For the pilot reliability study the subjects selected were within a restricted range, from PGY3 to PGY5, and did not fully represent the eventual target population. For the main study, the sample was more heterogeneous ranging from PGY1 to faculty. Even within the faculty there was a broad range of experience with almost all of the Women's College gynecology staff participating. Therefore, there was no over or under representation of any particular level of experience. Theoretically, one might still argue that the LSI is not applicable in a community setting because community gynecologists were not part of the sampling population. However, there is no reason
to expect that community based gynecologists are any less, or more, skillful than university faculty. Moreover, some of the fellows and final PG year participants were on the brink of becoming community gynecologists within the next several weeks to months and these individuals could reasonably be considered to be representative of the community gynecologists.

The "randomness" used in selecting raters for the pilot and main studies was, of course, theoretical in that we did not have access to the entire population of potential raters. However, it is unlikely that the process of rater selection would prevent generalizing LSI usage to raters in other institutions and communities. The study raters consisted of both university faculty and community raters. While rater A (the community rater) in the main study was clearly more lenient than the three faculty raters, the community rater in the pilot study did not perform differently from the university faculty raters.

Another aspect of generalizability relates to the item list. In particular, while *electrosurgical dissection* was hardly seen in the main study and was dropped from the analysis, it is a commonly used method of dissection in most community hospitals, where it serves as an alternative to laser dissection. Thus, the item should be retained in order to give the LSI sufficient flexibility to be usable in different settings.

Therefore, any methodologic limitation of subject and rater selection is theoretical only. Probably the most important factor in rater selection will not be whether the rater is university or community based, but rather, the quality of rater training.

10.4 Sensibility of the LSI

The sensibility of a clinical index is appraised with what might be called "enlightened common sense" which is a mixture of ordinary common sense plus a reasonable knowledge of
pathophysiology and clinical reality. Feinstein divided sensibility into five major topics and several subdivisions. The main topics are (1) purpose and framework (2) overt format (3) face validity (4) content validity and (5) ease of usage.

At this juncture the LSI is intended as an objective measure of level II laparoscopic skills in gynecology within the human setting, at a single point in time. If demonstrated to have construct validity (see Future research, 10.5), it could be used to assess a resident on several occasions during any three to six month gynecology rotation after which an aggregate score could be entered into the ITER. Once standards of performance are established through normative research (see Future research, 10.5) the index could also be used for credentialing gynecologists wishing to apply for laparoscopic privileges in both community and university settings in the situation where the applicant has trained in another country or in an environment where operative laparoscopy is infrequently practiced. This type of evaluation would occur during a preceptorship program.

The simplicity of usage of the LSI is evident in its additive index, numerical rating system, limited categories of response, clarity of the component elements and a scoring system in which the original expressions of data do not have to be converted by some form of transformation such as weighting. One aspect of sensibility is the replicability of an index which refers to the clarity and thoroughness of the directions provided for its usage. While the directions for rating have been facilitated through BARS, adjectival end anchors and a descriptor page, there appears to be some room for improvement. For example, the definition of electrosurgical dissection needs to be clarified so that raters are clear that applications of electrosurgical techniques for the sake of hemostasis or devitalizing a pedicle do not constitute electrosurgical dissection per se. Raters will require some direction about scoring trocar placements. For example, does a rater assign a score
based on the best or worst or average of the trocar insertion techniques? Finally, should initial inspection be scored at the initial point of the operation following laparoscope insertion or, subsequently, after various adhesions have been taken down to permit a more thorough inspection?

The LSI exhibits face validity. The nomenclature of the items is based on direct observations and the criteria defining each item is specified. The items appear to be sensible and reflect the attribute being measured.

Content validity was established by the panel of experts who participated in the development of the LSI. They agreed that the items were relevant and that they reflected the attribute to be measured. The items seemed comprehensive for the level of operative laparoscopy being assessed. Items which are not relevant or appropriate, such as instrument knowledge and suturing, have been eliminated. Therefore, there appear to be no inappropriate inclusions nor are there any obvious exclusions. In future, if the application of the LSI broadens to include level III gynecologic procedures and general surgery, it is likely that other items, such as suturing, intracorporeal knot tying and stapling will need to be incorporated into the index.

Finally, the LSI is characterized by feasibility and ease of usage. As mentioned previously, the scoring technique is simple. The time to complete an assessment after viewing a videotape or in vivo procedure requires less than five minutes, and there are no extra costs, including videorecording, which is standard procedure in many laparoscopy operating theatres.

10.5 Future research

As there are no previously developed measurements for assessing laparoscopic skills in humans, no gold standard exists and consequently criterion validity is not possible. An alternative is to establish construct validity which is an appraisal of the effectiveness with which an index does
its job in describing an existing or established construct. In the case of the LSI construct validity will be established by determining the effect of level of training on the global rating score and by convergent validity in which the LSI scores will be matched with the seven item global rating OSATS used in general surgery. Concurrent validity will be established by determining the level of correspondence between LSI scores and faculty rankings of postgraduate trainees, within their level of training.

A secondary study within the construct validity study will examine the agreement between LSI in vivo and videotape forms of administration. This is essential before videotape LSI assessments can be accepted as surrogates for in vivo assessments.

The purpose of normative research is to describe typical or standard values for characteristics of a given population. The importance of establishing standard values is axiomatic since the extent of deviation from these standards could potentially impact on the careers of aspiring gynecologic surgeons. The creation of data banks for laparoscopic surgery could prove helpful in the development of training systems and could provide a way for participants to know how successful their efforts have been at attaining a specified level of competence. Samples for normative studies must be large, random and representative of the populations heterogeneity and, consequently, the establishment of standard values is viewed as a long range project.

Responsiveness evaluates the extent to which a measure can detect true improvements or decrements over time. The more responsive an instrument, the more sensitive it is to clinically important changes, even when the changes are small. Several summary statistics describe the magnitude of change in those who experience a positive change. Most reflect a ratio of signal (observed change) over noise (some measure of variance). Studies should be undertaken to assess
the responsiveness of the LSI. For example, the magnitude of change from year I (the baseline year) to each succeeding level of training could be evaluated by calculating the standard effect size (SES) which measures the observed change divided by the standard deviation of the baseline score, or the standard response mean (SRM) which measures the observed change divided by the standard deviation of the differences. If the LSI is shown to be valid and responsive, it could subsequently be used as the metric in trials of interventions designed to improve laparoscopic surgical skills.

10.6 Conclusions

The LSI has been shown to be a sensible measure of level II gynecologic laparoscopic skill in the human setting. It has a high level of internal consistency and good reproducibility. In the short-term, construct validity and the level of agreement between in vivo and videotape administrations must be established. In the long-term, norms (standard values) should be developed, and responsiveness measured. The initial target population will be postgraduate residents in gynecology and staff gynecologists in community and university hospitals seeking privileges to perform operative laparoscopy. Ultimately, modified versions of the LSI could be developed for level III gynecologic laparoscopy and general surgery.
REFERENCES


Random House Dictionary of the English Language (Unabridged). Random House


Likert R. A technique for the measurement of attitudes. *Arch Psychol* 1932;21(140).


Guttman L. A basis for scaling quantitative data. *Am Sociol Rev* 1944;9: 139-150.


Appendix I: Guidelines for Resident Training in Endoscopic Procedures SOGC

This Committee Opinion has been prepared by the Junior Fellow Committee of the Society of Obstetricians and Gynaecologists of Canada and approved by its Council.

Operative endoscopy represents a rapidly increasing component of gynaecological surgery. It has progressively offered new approaches to old surgical problems. In many settings, operative endoscopy has set new standards for operative care. It has been in the environment of teaching centres that many of these new surgical methods have been proposed, sampled, and perfected. Recently, guidelines have been published by a sub-committee of the Society of Obstetricians and Gynaecologists of Canada (SOGC) regarding endoscopic training as it pertains to practising gynaecologists. At present, no such guidelines have been developed regarding training requirements for residents in obstetrics and gynaecology. The Junior Fellow Committee of the SOGC, representing all residents in obstetrics and gynaecology, assembled a sub-committee to address this issue and to suggest training guidelines in operative endoscopy as minimum requirements for current graduates in obstetrics and gynaecology.

Residency is essentially an apprenticeship with surgical training conducted under supervision. Similar to the preceptorship described in the SOGC guidelines for Endoscopic Surgery. By establishing standards of training in endoscopic surgery for all OB/GYN residency programmes, an expected level of competence in endoscopic procedures can be obtained among graduating trainees in obstetrics and gynaecology. It is hoped that these guidelines will facilitate appropriate utilization of endoscopic procedures and will provide better delivery of these services to patients.

The following guidelines are proposed pertaining to endoscopic training of residents in obstetrics and gynaecology.

1. As a minimum requirement, residents should be trained to competence in Level I and Level II Laparoscopic Procedures as outlined in Appendix I of the recent SOGC Policy Statement on Operative Endoscopy.

LEVEL I LAPAROSCOPIC PROCEDURES

- Laparoscopic sterilization
- Needle aspiration of simple cysts
- Ovarian biopsy
- Minor adhesions
- Laser/diathermy for endometriosis - revised AFS Stages I and II
- Linear salpingostomy and/or salpingectomy for ectopic pregnancy
LEVEL II LAPAROSCOPIC PROCEDURES

- Laparoscopic division of uterosacral ligaments
- Adhesiolysis for moderate and severe adhesions or adhesions involving bowel
- Salpingostomy for infertility
- Salpingectomy/salpingo-oophorectomy
- Laparoscopic ovarian cystectomy
- Endoscopic management of endometrioma
- Endoscopic surgery for revised AFS Stages III and IV endometriosis
- Laparoscopically assisted vaginal hysterectomy

2 Resident involvement in more advanced laparoscopic procedures (i.e. Level III procedures) should be strongly encouraged. When appropriate competence has been demonstrated on an individual basis, recognition in writing must be provided by the involved teacher or preceptor.

LEVEL III LAPAROSCOPIC PROCEDURES

- Myomectomy
- Pelvic lymphadenectomy
- Pelvic side-wall dissection
- Ureteral neurectomy
- Dissection of an obliterated Pouch of Douglas

3 Concurrent training in open operative procedures must be maintained. Competence in general gynaecological surgery is clearly a prerequisite for proficiency in any laparoscopic procedure. The trainee should demonstrate a full understanding of pelvic anatomy and the use of surgical instruments.

4 Appropriate training in the complications of endoscopic surgery and subsequent management should be provided. This training must include a full understanding of electrosurgical instruments and techniques, as well as methods for achieving haemostasis and indicators for laparotomy.

In conclusion, it must be stressed that the Junior Fellow Committee of the SOGC regards these guidelines as minimum training requirements for all current graduates in obstetrics and gynaecology. If such levels of training as outlined above cannot be provided by an OB/GYN training programme, then appropriate opportunities should be made available for residents to attain this level of training at other centres.

<table>
<thead>
<tr>
<th>Operative judgement</th>
<th>Intraoperative decision making unreliable with inappropriate use of consultation with staff.</th>
<th>Generally able to describe relations of operative care to clinical pictures but a) fails to consult with staff, and/or b) lacks confidence when making decision.</th>
<th>Able to relate operative care to clinical picture. Seeks appropriate advice when necessary.</th>
<th>Makes considered and appropriate intra-operative decisions. Appropriate consultation with staff.</th>
<th>Consistently makes considered, independent and appropriate intraoperative decisions. Deals appropriately with unexpected findings. Appropriate consultations</th>
</tr>
</thead>
</table>
Appendix 3: Information Sheet and Consent Form: Pilot Study

The Laparoscopic Skills Index: A Pilot Study

Investigators: J. Shime (416) 323-6126
R. Pittini

Purpose: The purpose of this study is to evaluate a new objective means of assessing technical skills at operative laparoscopy. The aim of this pilot study is to assess the amount of variation in scores assigned by a group of reviewers. The variation in scores between different reviewers and the variation in score to one review on different occasions will be examined in order to determine whether the evaluation method is consistent.

Procedure: Videolaparoscopy will be utilized to obtain videotape record of six operative laparoscopic procedures as performed by volunteer surgeons. These videotapes will be edited to highlight the relevant aspects with regards to the Laparoscopic Skills Index (LSI). The edited tapes will then be reviewed by six expert laparoscopists on two occasions and a score assigned using the LSI. The evaluations will be strictly anonymous with regard to both surgeon and patient. The emphasis will be on correlation of assigned scores rather than the actual scores.

Benefits Anticipated benefits are long term and include improved evaluation of current and future surgeons. Should this pilot study establish the LSI as a reliable means of evaluation a subsequent study of gynaecologists will be undertaken to determine validity of the scoring index.

Discomforts/potential Harms: Volunteer surgeons will not be asked to modify their surgical procedures in any form for the purpose of this study. As the LSI has yet to be validated and the credentials of the participating volunteers will have been established by other means, the scores assigned by the reviewers will be used only for the determination of interrater and intrarater reliability. The assigned scores will not be used in any fashion to directly evaluate those participating. All scores will be assigned in an anonymous fashion, with confidentiality assured. Only one co-investigator (RP) involved in editing of videotapes will be aware of the surgeon's identity. The tapes will remain in a teaching file and will not be labeled according to surgeon/patient.

Rights of Subjects: 1. The participation of surgeons is on a completely voluntary basis.
2. No repercussions for nonparticipation will occur.
3. All results will be presented in an anonymous fashion.
4. Strict confidentiality will be maintained for all participants.
CONSENT FORM

The Laparoscopic Skills Index: A Pilot Study

Investigators: J. Shime
R. Pittini

acknowledge that the research procedures described on the attached study information sheet, of which I have a copy, have been explained to me and that any questions that I have asked have been answered to my satisfaction. I have been informed that my participation is voluntary. The possible or potential harms have been explained to me. I know that I may ask now, or in the future, any questions I have about the study or research procedures. I have been assured that records (including videotape) related to my performance will be kept anonymous and confidential and that no information will be released or printed that would disclose my personal identity without my permission. No information obtained during the conduct of this study will be used for any purpose outside of the scope of the current outlined study.

I understand that I am free to withdraw from the study at any time. I further understand that if the study is not completed, or if there is withdrawal from it at any time, there will be no repercussions with regards to my status at Women's College Hospital.

I hereby consent to participate:

_________________________  _________________________
(signature of surgeon)        (name)

_________________________  _________________________
(signature of assistant)      (name)

_________________________  _________________________
(signature of witness)        (name)

The person who may be contacted about this research is: J. Shime who can be reached at: (416) 323-6126
Appendix 4: Rater Instruction Sheet: Pilot Study

Rater Instruction Sheet, LSI

You will be asked to view a videotape compilation of six separate operative laparoscopy procedures. The tape is divided into six sections, each section corresponding to one edited operative procedure 20 to 40 minutes in length. Each section will be clearly demarcated with a title. In each section there will be representative segments of the skills included in the Laparoscopic Skills Index (LSI).

You will be provided with six LSI reviewer forms to complete. One for each specific operative procedure. You will be allotted four weeks to complete the forms, however we ask that you complete viewing of any one procedure in the same sitting.

The LSI is designed to assess the technical skills used in operative laparoscopy.

The complete scale is comprised of 11 items. (an example form is included in the reviewer package) All 11 items may not be observed during a laparoscopic operation.
For this pilot study items 7, 8, and 11 (ligation, stapling and instrument knowledge) will not be used.

For the remaining eight items evaluated in this pilot study, computer forms will be used. Please TICK OFF your score, on a scale of 1 to 5, for each skill listed. TICK OFF number 6 if the skill was not observed. Do not tick off more than one number for each item.

NOTE: Do not attempt to provide a range in scores unless you feel it reflects the true range of skill observed. All scores may be the same if you feel this is an accurate reflection of the skills demonstrated.

A space will be provided on the form for written comments. Please feel free to offer comments on the difficulties encountered while scoring the segments.

For each criteria of the scale a descriptor of the important components of that item (skill) has been provided. Some descriptors, such as dissection, have multiple components. For a particular item, please score the item on the specific components observed. Some descriptors include criteria not amenable to videotape evaluation (e.g. Checks promontory before inserting trocar), please base your rating according to the remainder of the descriptor.

Please use the following guidelines in scoring each item:

1. POOR
2. 
3. AVERAGE
4. 
5. EXPERT
6. Not observed.

After assigning a grade for each item of the scale it is not necessary to calculate the final score. The completed forms will be collected at the end of the four week period.

The LSI form is part of a research study. The reviewer and the laparoscopists' names will be kept completely confidential.

Thank you for your assistance. If you have any questions, please contact Dr. J. Shime, Women's College Hospital (416) 323-6126.
Appendix 5: Item Descriptors Rater Information Package: Pilot Study

1. **Trocar Insertion**
   - checks promontory before inserting Verres needle
   - chooses appropriate site
   - avoids epigastric/others vessels
   - avoids viscera
   - keeps trocar tip in camera view during insertion.

2. **Initial Inspection**
   - performed in systemic, atraumatic fashion
   - identifies uterus, adnexae, ovarian formin, broad, round, and uterosacral ligaments, bladder fold, cul-de-sac, sigmoid, rectum, ilia, ureters, appendix, liver, gall bladder, diaphragm
   - identifies relevant pathology and plans surgical approach

3. **Laparoscope Control**
   - anatomic orientation
   - focuses
   - cleans lens with anti-fog/irrigation/against moist tissue
   - appropriate closeups and panoramas

4. **Suction/Irrigation**
   - balances suction and irrigation effectively
   - avoids trapping tissue in suction and atraumatically removes trapped tissue
   - clears smoke plume with minimal loss of pneumoperitoneum

5. **Dissection**
   - able to use laser, electrocautery, scissors, blunt forceps, or suction irrigation (+/- hydrodissection) to safely divide tissues, open tissue planes, delineates structures (eg vessels, ureter) and excise tissue
   
   i) laser - He Ne beam focused
      - chooses appropriate settings
      - always uses appropriate backstop
      - smooth motion, clear field (ie plume evacuated), good depth of penetration
   
   ii) electrosurgical - chooses appropriate settings
      - touch cut technique - activates electrode just before touching tissues
      - monopolar electrode well isolated and not touching adjacent instruments or tissue except area to be cut
   
   iii) blunt forceps - gently opens and closes without tearing vessels while cleaning away fat or areolar tissue or opening planes
   
   iv) irrigation probe - able to use for blunt dissection +/- hydrodissection

6. **Hemostasis**
   - identifies and isolates bleeders
   - uses bipolar for large vessels, venous sinuses and makes contact with both paddles
   - cauterizes with short bursts
   - uses monopolar for small bleeder during monopolar dissection
   - cautery tips are kept clean

7. **Ligation**
   i) Extracorporeal Tying
      - can perform extracorporeal knots (eg Roeder, Modified Weston, Square)
      - delivers knot firmly to desired site with pusher of endoknot technique
   
   ii) Intracorporeal tying
      - holds long end towards port through which tying end is passed
      - loop size adequate
      - when tying instrument is looped tip should be near to and facing short end
      - lays square knot with correct tension
   
   iii) Endoloop
      - backloads correctly
      - grasps pedicle after passing grasper through loop
      - dissects knot to desired site
      - ensures no other tissue included:
      - double cinches knot

8. **Stapling/Clipping**
   i) Stapler - appropriately loads and fires
      - chooses correct cartridge size
      - proper amount of tissue grasped
      - always has end of device in view before firing
      - checks for correct firing and loose staples
   
   ii) Clips - grasps only structure to be clipped
      - vessels should be skeletonized as clearly as possible
      - adequate space between clips if cutting to be done

9. **Tissue Removal**
   - chooses appropriate port size and location
   - able to dilate port if necessary
   - demonstrates safe morcellating techniques
   - if endopouch used:
      - backloads for insertion
      - irrigation to open pouch
      - delivers tissue into pouch
      - reduces tissue size in pouch
      - by cutting or suctioning
      - removes saline from pouch
      - before tightening knot to reduce diameter before extracting

10. **Final Inspection**
    - suctions fluid and clots from all quadrants
    - conducts underwater examination for bleeding
    - reinspects for bleeding after releasing pneumoperitoneum
    - instruments removed under visualization
    - releases pneumoperitoneum

11. **Instrument Knowledge**
    - knows connections and controls for insufflator, light source and camera
    - knows connections for suction/irrigator
    - knows connections for electrosurgical instruments and laser and chooses correct settings
    - makes reasonable decisions re trocar selection
Appendix 6: LSI Scoring Form: Pilot Study

**Laparoscopic Skills Index**

<table>
<thead>
<tr>
<th>Rater</th>
<th>Pilot Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Review #</td>
<td>1 2</td>
</tr>
</tbody>
</table>

**Criteria**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Poor</th>
<th>Average</th>
<th>Excellent</th>
<th>Not Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trocar Insertion</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Inspection of Abdomen</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Laparoscope Control</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Suction / Irrigation</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Dissection</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Hemostasis</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Tissue Removal</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Final Inspection</td>
<td>1 2 3 4 5</td>
<td></td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

**Comments**

Tape Identification □□□□□□

111
Appendix 7: LSI Rater Instruction Form: Main Study

The Laparoscopic Skills Index (LSI) is designed to assess a gynaecologic surgeons technical skills for performing operative laparoscopy. The LSI is composed of 13 items, each representing a skill which is employed during gynaecologic laparoscopy.

You will be asked to view a videotaped compilation of 20 separate gynaecologic laparoscopic procedures. This tape is divided into sections, each one corresponding to one operative procedure which has been edited to 30 minutes in length. Each operative procedure is clearly identified with a title. Each contains representative segments of skills listed in the LSI.

You will be provided with 20 LSI reviewer forms to complete, one for each operative procedure. While most of the items, such as trocar insertion and dissection, will be present in all procedures, some items such as knot tying/ligation, may not be. For each item (skill) observed please circle your score on a scale of 1 to 5 using the scoring descriptions underlining the numerical scale. If an item is not observed, circle the N/O. Do not circle more than one box for each item.

The following guidelines should assist you in assigning a score:

A score of 1 would describe the performance of a NOVICE, that is, an individual who continuously requires direction, performs the skill with hesitation and frequently requires correction.

A score of 5 would describe the performance of an EXPERT, that is, an individual who performs these skills smoothly in a controlled, progressive, self-directed, fashion.

The LSI is designed to assess surgical skills only. In assigning your score, please avoid being influenced by whether or not you feel the procedure was justified or appropriate.

It is important that you view each operative procedure in its entirety at one sitting. The entire compilation should be viewed over several days. Please do not review more than two procedures on any one day. Scoring for a particular operative procedure should be assigned immediately on completion of the viewing of that procedure. You will be allotted two weeks to complete and return the LSI forms.

The LSI form is part of a research study. Complete anonymity will be preserved for the patients, surgeons, and raters all of whom will be identified only by an assigned code number.

We are very grateful for your time commitment and assistance in this study. If you have any questions, please contact Dr. Jerry Shime, women's College Hospital (416-323-6126, Fax 416-323-6148, e-mail Shime@FTN.net).
Appendix 8: LSI Item Descriptors: Main Study

LSI ITEM DESCRIPTORS

Trocar Placement
- chooses appropriate sites
- chooses appropriate numbers
- keeps trocar tip in view throughout insertion
- avoids vessels (abdominal wall & intraabdominal)
- avoids viscera & adhesions
- avoids lunging
- parietal peritoneum is entered at right angle with minimal tenting of tissue

Initial Inspection
- systematic, atraumatic thorough (includes uterus, adnexae, pelvic side walls, ovarian fossae, cul-de-sac, uterovesical fold, uterosacral & broad ligaments, bladder, rectum, ureters, pelvic vessels, liver, gall bladder, diaphragm, appendix)

Laparoscope Control
- anatomic
- operative site centered on screen
- smooth motion
- appropriate use of zoom & panorama views

Irrigation
- uses irrigator purposefully to clear debris, hydrodissect & expose bleeders for coagulation.
It is not so excessive as to obscure field, interfere with coagulation or disperse debris.
- effectively balanced with suction

Suction
- clears fluid, plume & debris with minimal loss of pneumoperitoneum
- avoids tissue trapping
- atraumatically able to remove trapped tissue
- effectively balanced with irrigation

Maintenance of Visibility
- camera focus is maintained
- lens is cleared with irrigator or alcohol swab as necessary
- maintains an adequate pneumoperitoneum
- uses suction/irrigator effectively
- displaces or retracts obstructing structures effectively

Instrument & Tissue Handling
- uses appropriate instruments at all times, e.g.-nontoothed grasper on vascular or cystic structures
- safely, gently, meticulously & progressively divides tissues, opens tissue planes, delineates structures (eg vessels, ureter, cyst capsule ) & excises tissue
- avoids excessive stretch when applying traction e.g.-nontoothed graspers on vascular or cystic structures

Electrosurgical Dissection
- uses touch-cut technique, activating the electrode just prior to touching tissues
- able to use electrosurgical approach to safely divide tissues, open tissue planes, delineate structures & excise tissue
- uses appropriate power settings & cauterizes sufficiently to avoid bleeding but not excessively so as to cause thermal damage to nontargeted tissues
- monopolar electrode tip is kept in view at all times during activation
Laser Dissection
- HeNe beam is focused
- appropriately uses non-reflective backstop
- with a smooth motion & correct depth of penetration accurately directs laser to safely divide tissues, open tissue planes, coagulate small vessels & excise tissue

Knot Tying/Ligation
- able to perform extracorporeal knots (Roeder, modified Weston, square)
- able to deliver knot firmly to site with endoknot pusher
- intracorporeal typing:
  - holds long end towards port through which tying instrument is passed
  - places short end on side where knot is being formed
  - when tying instrument is looped its tape faces the short end for final knot formation
- Endoloop application:
  - backloads correctly
  - grasps pedicle after passing toothed grasper through loop
  - directs knot to desired site
  - ensures that only the desired pedicle is included
  - doubly cinches knot

Hemostasis
- uses bipolar for large vessels, oozing raw areas, venous sinuses & major pedicles
- irrigates first to expose bleedsers
- does not grasp too much tissue between bipolar paddles
- both paddles contact tissue to be cauterized
- when electrode tips stick to tissue they are removed with irrigation & short bursts of cautery
- uses monopolar or laser only for cauterizing small distinct vessels
- cautery tips are kept clean

Tissue Removal
- morcellation:
  - familiar with morcellation (apple coring technique)
  - keeps morcellator tips in view at all times
  - avoids injury to adjacent structures
- chooses appropriate port site & size
- if required able to correctly dilate port
- tissue is pulled through port via its narrow, longitudinal axis under continuous observation
- endopouch removal
  - backloads for insertion
  - irri tates pouch to open
  - delivers tissue into endopouch
  - reduces tissue size in pouch by cutting or decompressing as needed, suctions fluid from pouch before tightening knot to reduce diameter of contents

Procedure Completion
- inspects all quadrants
- maximally removes debris & fluid
- inspects operative site & cauterizes bleeding points
- also inspects for bleeding using underwater exam & partially decompressed abdomen
- all instruments are removed under direct observation including laparoscope
- trocar sites are checked for bleeding & managed appropriately
## Appendix 9: LSI Scoring Form: Main Study

**RATER:** (A-D)  
**PROCEDURE:** (1-20)  
**DATE COMPLETED:** (YY/MM/DD)

<table>
<thead>
<tr>
<th>Item</th>
<th>Novice</th>
<th>Expert</th>
<th>Not Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trocar Placement</strong></td>
<td></td>
<td></td>
<td>N/O</td>
</tr>
<tr>
<td>Uses inappropriate trocar number &amp; site selection; insertion technique risks injury to vessels or viscera</td>
<td>1 2 3 4 5</td>
<td>Ideal trocar number &amp; site selection; insertion technique is performed safely with trocar tip visible throughout</td>
<td></td>
</tr>
<tr>
<td><strong>Initial Inspection</strong></td>
<td>1 2 3 4 5</td>
<td>Systematic &amp; complete assessment of relevant pelvic &amp; abdominal structures</td>
<td>N/O</td>
</tr>
<tr>
<td>Unsystematic &amp; incomplete assessment of relevant pelvic &amp; abdominal structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Laparoscope Control</strong></td>
<td>1 2 3 4 5</td>
<td>Anatomic orientation &amp; centering are not maintained; fails to use zoom &amp; panorama effectively</td>
<td>N/O</td>
</tr>
<tr>
<td>Anatomic orientation &amp; centering are maintained, uses zoom &amp; panorama effectively</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Irrigation</strong></td>
<td>1 2 3 4 5</td>
<td>Irrigation is effectively used as needed to clear debris, hydrorrhcess &amp; coagulate bleeders</td>
<td>N/O</td>
</tr>
<tr>
<td>Erratic, excessive irrigation tends to obscure field, disperse debris or hinder coagulation of bleeders</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Suction</strong></td>
<td>1 2 3 4 5</td>
<td>Clears plume, fluid &amp; debris without loss of pneumoperitoneum; tissue trapping is rare &amp; releasedatraumatically</td>
<td>N/O</td>
</tr>
<tr>
<td>Fails to clear plume, fluid or debris adequately; tissue trapping &amp; loss of pneumoperitoneum occur often</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Maintenance of Visibility</strong></td>
<td>1 2 3 4 5</td>
<td>Keeps camera focused &amp; lens clear; effectively displaces or retracts structures to enhance view of operative site</td>
<td>N/O</td>
</tr>
<tr>
<td>Lens clearing &amp;/or camera focus are inadequate; has difficulty keeping operative site clear of intervening structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Electrosurgical Dissection</strong></td>
<td>1 2 3 4 5</td>
<td>Safely cauterizes &amp; divides tissues with minimal bleeding; keeps tips in view &amp; avoids non-targeted tissues</td>
<td>N/O</td>
</tr>
<tr>
<td>Cauterization often inadequate &amp; bleeding ensues or excessive with damage to adjacent or non-targeted tissues</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedure Completion</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>----------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Laser Dissection</td>
<td>Dissection is jerky or inaccurate; backstop use is poor &amp; plume control is inadequate</td>
<td>Dissection is smooth &amp; accurate; backstop use is appropriate &amp; plume control is excellent</td>
<td>N/O</td>
</tr>
<tr>
<td>Knot tying/Ligation</td>
<td>Knot tying or endoloop ligation is slow, awkward &amp; improperly placed or cinched</td>
<td>Quickly &amp; correctly performs intra &amp; extra corporeal knots &amp;/or endoloop ligation</td>
<td>N/O</td>
</tr>
<tr>
<td>Hemostasis</td>
<td>Difficulty exposing bleeders; poor choice of coagulation modes &amp; often coagulates excessively or inaccurately</td>
<td>Readily exposes bleeders; uses correct coagulation technique &amp; mode to obtain hemostasis safely &amp; effectively</td>
<td>N/O</td>
</tr>
<tr>
<td>Tissue Removal</td>
<td>Tissue morcellation or retrieval is awkward, unsafe or incomplete; poor choice of port for removal or poor port dilation technique</td>
<td>Uses correct technique to remove all tissue under direct vision through an appropriate port</td>
<td>N/O</td>
</tr>
<tr>
<td>Procedure Completion</td>
<td>Fluid, debris &amp; gas removal are incomplete; does not inspect for bleeding; does not observe instrument removal</td>
<td>Fluid, debris &amp; gas removal are complete; thoroughly inspects for bleeding; all instruments removed under observation</td>
<td>N/O</td>
</tr>
</tbody>
</table>

Please feel free to comment on any difficulties encountered while scoring a procedure:
In gynecology, laparoscopic surgery has increased dramatically over the past decade. The potential benefits of this type of surgery include smaller incisions, less blood loss, fewer postoperative symptoms and medication requirements, a shorter length of hospital stay, and earlier return to normal activities. Sometimes these benefits are offset by unfavourable outcomes which, at least in part, relate to limited experience in operative laparoscopy and unstandardized credentialing guidelines.

The Laparoscopic Skills Index (LSI) is a new measurement tool designed to test the skills of gynecologists in the performance of operative laparoscopy. Before the LSI can be put into use it first must be shown to be reliable (that is, repeated assessments by different raters are in agreement) and valid (that is, it must be shown to truly be a measurement of operative laparoscopy skills).

The investigators named above have undertaken the task of assessing the reliability and validity of this new measurement tool. To do so, we require videotapes of several operative laparoscopic procedures which will enable us to demonstrate the various skills required to perform this type of surgery. Ordinarily laparoscopy is performed by the surgeon viewing intra-abdominal structures on a video monitor similar to a television screen. Like an ordinary television set, the monitor is attached to a VCR machine so that videotaping is easy to accomplish. The videotape recording shows only what is taking place inside of the abdomen. No identifying features including the patient, surgeon, nurse or equipment are visible on the videotape.

Because you and your gynecologist have decided to proceed with operative laparoscopy we are seeking your permission to videotape your surgery. Videotaping of operative laparoscopy is commonly done and it in no way affects the manner in which surgery is performed or the surgical outcome. The videotape produced will be strictly anonymous and both you and your surgeon will be identifiable only by a code number known to one of the investigators (RP). All videotapes will remain in a teaching file and will not be labeled with the names of the patients, surgeons assistants, or dates of surgery.

**Benefits/Harms**

There will be no direct benefit or harm to you as a result of your participation in this study. The benefits will result from the potential application of the LSI as a measure of laparoscopic surgical skills.
Rights of Subjects

1. The participation of patients is on a completely voluntary basis.
2. The participation of surgeons is on a completely voluntary basis.
3. No repercussions for nonparticipation will occur.
4. All videotapes and data will be stored in an anonymous fashion.
5. Strict confidentiality will be maintained for all participants.
PURPOSE: The purpose of this study is to determine the reliability of a new, objective measure of assessing surgical skill at gynecologic operative laparoscopy.

The Laparoscopic Skills Index (LSI) is a multi-item index which uses a five point behaviourally-anchored rating scale for each item. The index was designed by a consensus technique to assess technical skills at a single point in time and, therefore, is a status index.

RESEARCH METHOD: Videolaparoscopy will be utilized to obtain videotaped records of 20 gynaecologic laparoscopic procedures performed by volunteer surgeons and residents. Videotapes will be edited to highlight their relevant aspects of the LSI. Edited tapes will be reviewed by four expert laparoscopists and a score assigned for each operative procedure. All assessments will be strictly anonymous with regard to the surgeon and patient who will be identified only by code number. The reviewers (raters) will also be identifiable only by code number. Emphasis will be on correlation and concordance of assigned scores rather than actual scores.

BENEFITS: Objective structured assessment of technical skills (OSATS) are becoming increasingly important in the evaluation process of residents in surgical specialties. This is the first laparoscopic skills assessment form in gynecology. If the LSI is shown to be reliable, valid, and easy to use, it will permit objective evaluation of operative laparoscopic skills in residents performing gynecologic laparoscopy. Other potential benefits could include: extension of the LSI to general surgical residents, and gynecologists seeking privileges to perform gynecologic laparoscopy. Finally, the LSI has the potential to serve as a measure of the effect of skills instruction programmes in laparoscopic surgery.

DISCOMFORTS/POTENTIAL HARMs: Volunteer surgeons will not be asked to modify their surgical procedures in any form for the purpose of this study. As the LSI has yet to be validated and the credentials of the participating volunteers will have already been established by other means, the scores assigned by reviewers will only be used for the determination of inter-rater reliability. The assigned scores will not be used in any fashion to directly evaluate study participants. All scores will be assigned in an anonymous fashion with confidentiality assured. Only one co-investigator involved in editing of the videotapes will be aware of the surgeon's identifies. The tapes will remain in a teaching file and will not be labelled according to surgeons/patients.

RIGHTS OF SUBJECTS:
1. The participation of patients is on a completely voluntary basis.
2. The participation of surgeons is on a completely voluntary basis.
3. No repercussions for nonparticipation will occur.
4. All videotapes and data will be presented in an anonymous fashion.
5. Strict confidentiality will be maintained for all participants.
SURGEON CONSENT FORM

THE LAPAROSCOPIC SKILLS INDEX: A RELIABILITY STUDY

Investigators: J. Shime (416-323-6126)
R. Pittini

I acknowledge and understand the research procedures described on the attached study information sheet of which I have a copy. Any questions that I have asked have been answered to my satisfaction. I am aware that my participation is voluntary. The possible or potential harms have been explained to me. I know that I may ask now, or in the future, any questions which I have pertaining to the study or research procedures. I have been assured that the records (including videotapes) related to my performance will be kept anonymous and confidential and that no information will be released or printed that would disclose my personal identity without my permission. No information obtained during the conduct of this study will be used for any purpose outside the scope of the research study.

I understand that I am free to withdraw from the study at any time. I further understand that if the study is not completed, or if there is withdrawal from it at any time, there will be no repercussions with regards to my status at Women's College Hospital.

I hereby consent to participate.

__________________________  _______________________
(signature of surgeon)        (name)

__________________________  _______________________
(signature of assistant)       (name)

__________________________  _______________________
(signature of witness)         (date)

The person who may be contacted about this research is Dr. J. Shime (416) 323-6126
PATIENT CONSENT FORM

THE LAPAROSCOPIC SKILLS INDEX: A RELIABILITY STUDY

Investigators: J. Shime (416-323-6126)
               R. Pittini

I acknowledge that the research procedures described on the attached study information sheet, of which I have a copy, have been explained to me and that any questions that I have asked have been answered to my satisfaction. I have been informed that my participation is voluntary. The possible or potential harms have been explained to me. I know that I may ask now, or in the future, any questions I have about the study or research procedures. I have been assured that records (including videotapes) related to my surgery will be kept anonymous and confidential and that no information will be released or printed that would disclose my personal identity without my permission. No information obtained during the conduct of this study will be used for any purpose outside the scope of the outlined study.

I understand that I am free to withdraw from the study at any time. I further understand that if the study is not completed, or if there is withdrawal from it at any time, and that my withdrawal will, in no way, affect my treatment.

I hereby consent to participate.

_________________________________________   ________________________________
(signature of patient)                       (name)

_________________________________________   ________________________________
(signature of witness)                       (name)

_________________________   ____________________
(date)

The person who may be contacted about this research is Dr. J. Shime (416) 323-6126
June 9, 1999

Dr. J. Shime
Department of Obstetrics & Gynaecology
Women's College Hospital

RE: App #9661 - “The Laparoscopy Skills Index: A Pilot Study”
Investigators: Pittini R, Shime J

Dear Dr. Shime,

Thank you for submitting amendments to the Study Information Sheet for Patients for the above study (APP#9661). The amendments have been reviewed and the study has been reapproved by the Research Ethics Board, Women's College Campus, for one year.

Best wishes on the successful completion of your research

Yours truly,

Russ Springate, MD, CCFP
Chair, Research Ethics Board
Women's College Campus
Appendix 12: Memorandum to staff re planned LSI Main Reliability Study

MEMORANDUM

TO: Ob/Gyn Staff

FROM: Jerry Shime and Richard Pittini

DATE: June 2nd, 1999

RE: Laparoscopic Skills Index

Many of you know that we are working on a new measurement of laparoscopic surgical skills, the Laparoscopic Skills Index. Some of you contributed suggestions for items to be incorporated, others provided us with videotapes and yet others served as videotape reviewers during a previous pilot study of the LSI.

To complete the reliability study of the LSI we will need several videotapes of laparoscopic procedures performed by residents and members of the staff over the next several weeks. We would actually like to obtain 25 to 30 cases. Ideally, we would like procedures with varying degrees of difficulty and performed by surgeons or residents. Each case will have to be taped from beginning to end and one of us (RP) will edit the tapes. The purpose of accumulating these tapes is to have demonstrations of the various items which comprise the laparoscopic skills index. We will also excerpt components of the tapes for the purpose of developing an instructional video.

One of us (RP) and Dr. Veena Zaveri, the endoscopy fellow, will be responsible for actually inserting the videocassette and starting the videorecording and collecting the tape.

We will require that both you and your patient sign a consent to permit videotaping. Sample consent forms are appended but we will send you a supply of the official forms later this week. Please note that all videorecordings will be anonymous and identifiable by a code number known to only one of the investigators (RP) and not the other (JS). Please notify Dr. Zaveri through my office (323-6126) when you have a case eligible for videotaping.

It is essential that we complete the videotaping by late July at the latest and we would therefore be extremely grateful if each of you could take the time to provide us with three to four cases over the next six to seven weeks. Thank you very much for your assistance with this study.

JS jd

123
Appendix 13: Individual Rater Trend Lines for Total Scores (percent) Assigned to Each Subject: Main Reliability Study