Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale

CR Larsen,a T Grantcharov,b L Schouenborg,a C Ottosen,a JL Soerensen,a B Ottesena

a Department of Gynaecology and Obstetrics, The Juliane Marie Centre (for Children, Women and Reproduction), Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark b Department of Surgery, University of Toronto, St Michael’s Hospital, Toronto, Ontario, Canada

Correspondence: Dr CR Larsen, Department of Gynaecology, The Juliane Marie Centre (for Children, Women and Reproduction), Copenhagen University Hospital, Rigshospitalet, Blegdamsvej 3, DK-2100 Ø, Copenhagen, Denmark. Email crl@rh.regionh.dk

Accepted 27 February 2008.

Objective The purpose of this study was to develop a global- and a procedure-specific rating scale based on a well-validated generic model (objective structured assessment of technical skills) for assessment of technical skills in laparoscopic gynaecology. Furthermore, we aimed to investigate the construct validity and the interrater agreement (IRA) of the rating scale. We investigated both the gamma coefficient (Kendall’s rank correlation), which is a measure of the strength of dependence between observations, and the kappa value for each of the ten individual items included in the rating scale.

Design Prospective cohort, observer-blinded study.

Setting Departments of Obstetrics and Gynaecology in Zealand, Denmark.

Population Twenty one gynaecologists or gynaecological trainees.

Material and methods Twenty-one video recordings of right side laparoscopic salpingectomies were collected prospectively, eight from novices (defined as <10 procedures), seven from intermediate experienced (20–50 procedures) and six from experts (>200 procedures). All operations were performed by the same operative principles and using a standardised technique. The recordings were analysed by two independent, blinded observers.

Main outcome measures Construct validity of the rating scale based on operative performance (median of total score) and interrater reliability.

Results There were significant differences between the three groups: median score of novices 24.00 versus intermediate 29.50 versus expert 39.50, $P < 0.003$) The IRA was 0.83 overall. The gamma correlation coefficient was 0.91. The kappa values varied from 0.510–0.933 for each of the individual items of the rating scale.

Conclusions The procedure-specific rating scale for laparoscopic salpingectomy is a valid and reliable tool for assessment of technical skills in gynaecological laparoscopy.

Keywords Assessment, construct validity, gynaecology, interrater reliability, laparoscopy, salpingectomy.

Introduction

Laparoscopy, training and assessment

An increasing number of gynaecological surgical procedures are presently performed by laparoscopic technique. This leads to an increasing demand for evidence- and proficiency-based education, training and assessment of laparoscopic skills. Traditionally, education has been based on the apprenticeship model leaving both training and assessment of knowledge as well as technical skills subjective and unstructured.1 So far, little has been carried out to develop and integrate structured basic training in laparoscopy in the surgical curriculum, and no validated structured system for objective assessment of technical surgical skills in gynaecological laparoscopy is available. Consequently, we need a feasible, structured and objective system for assessment of both technical and procedural skills.

In Denmark, 94% of all surgically treated ectopic pregnancies are managed by laparoscopy.2 The laparoscopic salpingectomy was in the present study chosen as the operative procedure for developing a method for objective structured assessment of technical surgical skills based on human operations. Laparoscopic salpingectomy is a basic laparoscopic procedure possessing the necessary complexity for skills
assessment. Furthermore, salpingectomy is one of the first laparoscopic procedures a trainee is exposed to. Finally, all specialised gynaecologists and obstetricians working on call should be able to perform a laparoscopic salpingectomy.

Assessment systems
Several systems to assess technical skills in minimally invasive procedures have been developed; some based on error detection and some on rating the technical skills. The best described error detection systems is the Time–Error matrices by Seymour et al. and the Observational Clinical Human Reliability Assessment system (OCHRA) by Tang et al. The Time–Error matrix has later successfully been modified to a more detailed version, closer to the very complex OCHRA. Characteristic for both systems is that they are highly detailed, thereby very time-consuming and difficult to implement and use for the observers, resulting in low feasibility.

Objective structured assessment of technical skills
In the late 1990s, Martin et al. and Reznick et al. developed the approach for assessing technical skills called Objective Structured Assessment of Technical Skills (OSATS). The OSATS was developed on the basis of the Objective Structured Clinical Examination. OSATS was originally developed for bench station test and consists of a task-specific checklist and global rating scale (GRS). The GRS has seven items, each evaluated on a global 5-points Likert-like scale where the lowest, middle and highest scores are defined by explicit descriptions of performances.

It is well documented that GRSS are reliable, have high interrater reliability (IRR) and construct validity. Since first presented, the OSATS has been modified and tested in many different surgical areas such as open surgery, laparoscopic surgery, vascular surgery, microsurgery, urology, ophthalmology (Global Rating Assessment of Skills in Intracocular Surgery), gynaecology and obstetrics. Most tests were conducted using OSATS as a bench test, fewer in clinical set-up, and none of them for laparoscopic gynaecology. In the different studies referred to above, the Cronbach’s coefficient alpha (expressing the internal consistency reliability) varies form 0.719 to 0.9721. Unfortunately, IRR (interrater agreement [IRA]), which expresses the proportion of times to which two or more independent observers agree absolutely on their rating of a subjects performance, is not always described in these studies. An overall agreement among two observers ≥0.8 is, based on expert opinion, considered acceptable for a test system. In those studies where the IRR is stated, it varies from 0.709 to 0.9723. The solid evidence of the reliability, feasibility and construct validity of OSATS, and the modified versions for different specialties, have almost completed the OSATS as the gold standard in assessment of technical skills. The Toronto group creating the OSATS also made a modified assessment system for evaluation video recordings of laparoscopic (gastrointestinal) surgery. The modified system was developed for operations on anaesthetised pigs, not human operations. The system consists of a reduced GRS suitable for laparoscopic surgery and a task-specific rating scale called operative component rating scale (OCRS).

Objective
The purpose of this study was to develop, and investigate the construct validity, IRA and gamma correlation on a procedure-specific rating scale for objective structured assessment of technical surgical skills in laparoscopic salpingectomy.

Methods
After a hierarchical task analysis of the laparoscopic salpingectomy, we constructed a modified rating scale for laparoscopic salpingectomy called Objective Structured Assessment of Laparoscopic Salpingectomy (OSA-LS; Table 1). The OSA-LS was based on both the original OSATS and the modified rating scale for laparoscopic cholecystectomy, developed by Grantcharov et al. It consists of five general items and five task-specific items equivalent to the OCRS by Dath et al. First, in a pilot study, the observers together assessed ten video recorded operations to standardise their assessments and to adjust the scale (data not shown). In the main study, 21 video recordings of right side laparoscopic salpingectomies were collected prospectively over a 6-month period, 8 performed by novices (defined by less than 10 procedures), 7 by intermediate experienced (20–50 procedures) and 6 by experts (>200 procedures). All the recorded operations were performed by different surgeons, but using a standardised operative technique (based on expert consensus) as described by Nezhat et al. (Table 2). The two independent observers used the OSA-LS chart for assessment of the 21 unedited video recordings. The observers (L.S. and C.O.) were blinded for surgeon and proficiency group status. Both observers are experts in laparoscopic gynaecological surgery, having performed more than 2000 advanced laparoscopic procedures each. The introduction of Veress needle and placement of the trocars were not evaluated in this study.

Ethics
No additional ethical approval was needed according to the Danish National Committee on Biomedical Research Ethics.

Statistics
Data on the performance on each of the individual items included in the rating scale are ordinal. Ordinal data are categorical data where there is a logical ordering of the categories. The Likert-like scales that are also used in many surveys.
is a typical example: 1 = strongly disagree; 2 = disagree; 3 = neutral; 4 = agree; 5 = strongly agree. Cumulated scores for subjects or groups of subjects are continuous data. Due to the sample size and the nature of the results, a Gaussian distribution could not be expected. Cumulated scores are presented as median and interquartile range (IQR) and compared using the Kruskall–Wallis nonparametrical comparison of mean. For post hoc analysis, Bonferroni corrected Mann–Whitney U tests were used. A P value of (two-tailed) <0.05 is considered to be statistically significant. IRA was calculated as observation events agreements divided by the total number of observations for the single proficiency groups as well as for the entire sample. The gamma coefficient, a nonparametric rank correlation investigating agreement in ordinal categorical data, was used to investigate strength of correlations among the observers at single subject level. Systematic as well as nonsystematic disagreements were also analysed. Finally, to reveal items less agreeable, kappa values on single items level
are calculated. Analysis was performed using the Statistical Package for Social Sciences (SPSS \textsuperscript{13.0} for Windows (SPSS Inc., Chicago, IL, USA). Graphics was made on Graphpad Prism 4.0 for windows (Graphpad Software Inc., San Diego, CA, USA).

### Results

The independent and blinded evaluation of the operations demonstrated that the median score in the novice group was 24.00 (IQR 23.75–25.25), in the intermediate experienced group was 29.50 (IQR 28.00–31.00) and in the expert group was 39.50 (IQR 33.50–42.50). This revealed that the OSA-LS was construct valid and able to discriminate between all groups ($P < 0.03$) (Figure 1 and Table 3). The difference in overall score between novices and intermediate experienced gynaecological surgeons was 6 points and between intermediate and experts 8 points. The overall IRA was 0.831, varying from 0.759 in the experienced group to 0.905 among the intermediate experienced gynaecological surgeons (Table 4).

### Discussion

#### Construct and discriminative validity

The OSA-LS for video evaluation of surgical skills in laparoscopic gynaecology was demonstrated to be feasible, had good construct validity and high IRA. Construct validity, a core property of a test, is the extent to which the test measures the trait that it purports to measure. The OSA-LS for video...
assessment in the operating room demonstrates construct validity, like the bench station OSATS did. Our results are consistent with the results found in other clinical specialties.9,10,12–14,16–18 The groups with different levels of experience were clearly discriminated by both observers using the OSA-LS. Based on these findings, it can be concluded that the test can be applied to test the laparoscopic abilities of trainees in obstetrics and gynaecology.

A major advantage of this OSA-LS system is that the assessment is based on a real human operation rather than a bench-simulator- or animal model, thereby testing how the surgeon is actually performing in operating room on humans rather than how they intend to perform the surgery by demonstrating the procedure in a model, simulator or animal.

According to the classic Miller’s pyramid of competence development,27 the ‘Level 3: shows how’ can be represented by evaluation of competence in the bench station test or simulator, whether the ‘Level 4: Does’ only can be represented by evaluation of competence in a real (human) operation. The OSA-LS provides us ability to test the competence at the ‘Level 4’.

Another advantage of the OSA-LS is that it can serve as an excellent basis for structured feedback. Going through the operation video together with the OSA-LS evaluation could provide the trainee with valuable knowledge of his or her strengths and weaknesses and can potentially shorten the learning curve.28

Nevertheless, the sample size in the present study is small, as in most of this kind of studies,9,20,21 consequently the results have to be interpreted with caution.

### Performance range within the groups

Ideally, the range of performance should be narrow within each group, indicating that the subjects fit the group definition. In this study, however, we found a quite wide performance range in the expert group and a narrow performance range in the novice group. There could be several explanations for this. First, the figures could be coincidental, due to small sample size. More likely, the wide performance range could indicate that some experts are more skilled than others. The definition of the proficiency groups is traditionally

### Table 3. Construct (discriminative) validity of the OSA-LS: Comparison of median score of the three proficiency groups. Statistics: Kruskall–Wallis nonparametric, post hoc analysis: Mann–Whitney (Bonferroni corrected)

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Median</th>
<th>25% percentile</th>
<th>IQR</th>
<th>75% percentile</th>
<th>Range</th>
<th>Kruskall–Wallis P value</th>
<th>Mann–Whitney (post hoc) P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>8</td>
<td>24.00</td>
<td>23.75</td>
<td>25.25</td>
<td>25.50</td>
<td>25.00</td>
<td>23.50</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Intermediate</td>
<td>7</td>
<td>29.50</td>
<td>28.00</td>
<td>31.00</td>
<td>33.00</td>
<td>28.00</td>
<td>Intermediate vs experts &lt;0.03</td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td>6</td>
<td>39.50</td>
<td>33.50</td>
<td>42.50</td>
<td>44.00</td>
<td>31.50</td>
<td>All</td>
<td>—</td>
</tr>
</tbody>
</table>

### Table 4. Inter Rater Reliability at proficiency group level and at the overall level (bold)

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Items</th>
<th>Numbers of evaluated observations</th>
<th>Number of disagreement</th>
<th>IRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>8</td>
<td>9</td>
<td>72 (9 × 8)</td>
<td>14</td>
<td>0.806</td>
</tr>
<tr>
<td>Intermediate</td>
<td>7</td>
<td>9</td>
<td>63 (9 × 7)</td>
<td>6</td>
<td>0.905</td>
</tr>
<tr>
<td>Expert</td>
<td>6</td>
<td>9</td>
<td>54 (9 × 6)</td>
<td>13</td>
<td>0.795</td>
</tr>
<tr>
<td>Overall</td>
<td>21</td>
<td>9</td>
<td>189 (9 × 21)</td>
<td>32</td>
<td>0.831</td>
</tr>
</tbody>
</table>

Item 10 excluded, see text for details.
Table 5. Kappa values on single items level

<table>
<thead>
<tr>
<th>Item</th>
<th>n</th>
<th>Kappa</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower bound</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>0.851</td>
<td>0.100</td>
<td>0.655</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>0.653</td>
<td>0.132</td>
<td>0.394</td>
</tr>
<tr>
<td>3</td>
<td>21</td>
<td>0.933</td>
<td>0.065</td>
<td>0.806</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>0.625</td>
<td>0.129</td>
<td>0.372</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>0.789</td>
<td>0.112</td>
<td>0.569</td>
</tr>
<tr>
<td>6</td>
<td>21</td>
<td>0.510</td>
<td>0.149</td>
<td>0.218</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>0.731</td>
<td>0.147</td>
<td>0.443</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>0.695</td>
<td>0.137</td>
<td>0.426</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>0.695</td>
<td>0.131</td>
<td>0.438</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Objective assessment of surgical competence in laparoscopy

carried out by number of surgical procedures performed or times spent in a given position. This is not the most appropriate way as the number of cases performed is not an objective measure of competency. Furthermore, individuals have different learning curves. Some individuals have innate abilities to perform laparoscopic surgery and a very steep learning curve, others may need a variable number of procedures to reach a plateau and some may never achieve proficiency due to poor neuropsychological abilities. In fact, only the novices represent a truly defined proficiency group, quantitative as well as qualitative; having performed none or few procedures and being at the early stage of their learning curve. It is more difficult to define the intermediate and expert groups by a quantitative definition, such as number of procedures performed or at the early stage of their learning curve. When a person has reached a plateau, it might be needed to obtain valid information. This observation is however, must be developed and validated in a separate prospective study.

Re-analysing the recordings used in this project, we believe that there are not only differences in level of difficulty of a given procedure, but there are also substantial differences in the performance levels among the experts. Additionally, compared with the novices, who all handled the tissue extremely gently, some of the experts seemed confident and fast, handling the tissue a bit roughly. This might not influence the surgical outcome but is detected in the OSA-LS system.

Talent selection

A continuing discussion among health authorities is whether assessment systems in surgery can be a method for recruitment and career guidance. Based on the figures from this study, and based on previous studies on simulator-based salpingectomy, this seems to be quite difficult. When the range is as narrow as seen in this investigation, it would be extremely difficult to distinguish talents from non-talents, unless more video recordings per individual were evaluated. It is, however, unknown how many video recordings would be needed to obtain valid information. This observation is consistent with findings in the original paper where it was stated that OSATS were not designed as a predictor of surgical skills for residents before entering specialty training.

Interrater agreement

The level of agreement between two independent observers blinded for the test subjects training status is important in the evaluation of an assessment system. It reveals how unambiguous the test is, and thereby how valid it is if used by different independent raters. Several methods establishing the IRA are presented in the literature. Cohen’s kappa value is often described as the best measurement for the degree of agreement among the observers. Fundamentally, kappa calculates the degree of agreement, but it also takes into account the degree of agreement that could be expected to occur by chance, hence argued to make this statistical test more robust (beyond-chance agreement). This is very important in a single item test, and when the outcome is binary, e.g. yes or no, where the agreement occurring by chance can be as high as 25%. In a multiple items test, like this modified ten items OSA-LS, evaluated on a five categories Likert-like scale, the
overall agreement occurring by chance is negligible. This makes the simple calculation of IRA (observation events agreements/total number of observation) a sufficient measure for general purpose. A disadvantage of using only kappa statistics (and IRA) is that kappa only gives a general value of the observer agreement. It does not make distinctions in-between various types and sources of disagreement. Besides, in a multi-item test like the OSA-LS presented, an IRA > 0.8 is perfectly acceptable but still leaves us with 20% disagreement and no information on where to find the disagreement. Consequently, we also investigated the correlations coefficient gamma and the kappa values at a single items level.

**Gamma coefficient and kappa value**

The gamma coefficient is a nonparametric rank correlation investigating agreement of ordinal categorical data. The gamma coefficient was used to investigate strength of correlations, or agreement, among the observers at single subject level. This test can reveal whether a possible disagreement is systematic or random among the observers, or if possible, disagreement is only found in certain individuals or groups of individuals. Values of the gamma coefficient range from –1, negative association to +1, perfect agreement; 0 indicates absence of association. To explore that, we calculated the correlation of the observers to see whether a small discrepancy was systematic or nonsystematic and in which groups the correlation was highest. Figure 1 shows that the novice group subjects are very homogeneous in total score. However, this group also demonstrated the slightest correlation among the observers. In contrast, this disagreement is nonsystematic, leaving this group with the smallest dispersion. In this study, we found a very high overall correlation among the observers. As seen in the scatter plot, the correlation was higher for the intermediate and expert groups than for the novice group, but the discrepancy found in the latter was not systematic and thereby not influencing the total OSA-LS score for the individual subject. Based on these results, we conclude that the OSA-LS is suitable regarding correlation and systematic as well as random disagreement.

Looking at single items level, kappa values discovered some items more disagreeable than others. Items 2, 4 and 6 had the lowest kappa values, revealing quite some disagreement among the observers. This information could be used to discuss the interpretation of the item terms, and if necessary specify or rephrase the terms. If there is still a low degree of agreement, it should be considered to exclude the item from the list. However, in this study, the confidence intervals are large, most, except item 3, suggest that the level of agreement ranges from poor to excellent for each item making conclusions on kappa values difficult in a study of this size. Furthermore, it demonstrates that the simple IRA calculation probably is more suitable for this kind of observations, Likert-like scales with a small sample size.

**Strengths and weaknesses using the rating scales**

In contrary to the traditional method known from the apprenticeship educational model, the OSA-LS for laparoscopic gynaecology provides a valid, structured, objective and systematic method for assessment and feedback of technical skills. Besides valid and detailed assessment of the trainee, the global- and task-specific rating scales also provide clinical relevant information for constructive feedback. This is a great advantage compared with the use of laparoscopic simulators, box trainers and other metrics based systems for evaluation of skills. The simulator method although, is construct valid, but the metrics used in the assessment, like instrument angular path and instrument path length, have only relevance for improvement of dexterity, they do not have direct clinical implication. These parameters do therefore not provide clinically useful information for feedback on operative technique. Using the rating scales, the assessor can feedback the marks given for the different items to the trainee while going through the video recording, providing examples on where to improve. This could be carried out in a nonstressful setting and provide structured, objective and very detailed advice.

A disadvantage of the OSA-LS video evaluation is how time-consuming it is. However, it is an advantage that the evaluation is video based, giving the assessor the possibility to analyse the performance whenever it is convenient time wise. The fact that the evaluation can be blinded for surgeon is also a big advantage for objective assessment. The study has a possible bias by using the same experts for development of the OSA-LS scale and for the later validation of the rating scale. The internal consistency is high, based on the high IRR, while the external consistency has to be demonstrated by applying this rating scale in other departments and with other raters.

**Certification**

In certification matters, it is extremely important that a test is valid. Refusing a trainee with sufficient skills to operate will be inconvenient, to let a trainee surgeon with inadequate skills operate would be unacceptable. The high degree of construct validity and IRA makes it possible that this rating scale can serve as a basis in high stakes situations like certification and recertification. However, this study was not designed to define ‘cut off level’, this will still have to be based on expert consensus according to national surgical curricula.

**Learning curves**

Developing procedure-specific rating scales for different procedures also provide a good opportunity for assessment of learning curves for the different procedures. The learning curves combined with the advantages of feedback using the rating scales open the possibility to design high-quality training curricula in advanced laparoscopy.
Future studies
Using this validated OSA-LS scale for laparoscopic salpingectomy is of obvious interest to test the impact of simulator training of laparoscopic skills on real operations. Further investigations on the learning curve of individual trainees will also be of great interest. Finally, a second modification on the scale including a difficulty grading system to be applied on nonstandard cases should be developed.

Conclusion
The OSA-LS for assessment of technical surgical skills in laparoscopic gynaecology is construct and discriminative valid and has a sufficiently high degree of IRA and gamma correlation among independent observers blinded to surgeon. The kappa values at single item level revealed that seven of ten items represented a high IRA; only items 2, 4 and 6 needed better definitions or more training of the observers. The system can provide the trainee both objective assessment and detailed clinical relevant feedback.

Funding
Rigshospitalet, University Hospital of Copenhagen, funded the project.

Details of ethics approval
The Helsinki II declaration as well as local legalisations were respected, no additional ethical approval was needed according to the Danish National Committee on Biomedical Research Ethics; the video recordings were made anonymous, patients cannot be identified or tracked.

Contribution to authorship
C.R.L.: Principle investigator, design, acquisition of data, statistics, analysis and interpretation of data and drafting the paper. T.G.: Background research, design, results analysis and discussion and revising the paper. L.S.: Design, data collection and analysis and revising the paper. C.O.: Design, data collection and analysis and revising the paper. J.L.S.: Background research, design, statistics, discussion analysis and revising the paper. B.O: Supervisor, fundraiser, design, results collection and analysis and revising the paper. J.L.S.: Background research, design, statistics, discussion analysis and revising the paper. C.O.: Design, data collection and analysis and revising the paper. T.G.: Background research, design, results analysis and discussion and revising the paper.

Acknowledgements
Thanks to Svend Kreiner, Associate professor at Department of Biostatistics, Faculty of Health Sciences University of Copenhagen, for statistical support. Thanks to the departments of Gynaecology and Obstetrics at the Hospitals in Gentofte, Herlev, Hillerød, Hvidovre, and Roskilde for collecting video recordings of operations.

References


