DOI: 10.1111/j.1471-0528.2008.01732.x
www.blackwellpublishing.com/bjog

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 Ottesen[a]

[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.

Please cite this paper as: Larsen C, Grantcharov T, Schouenborg L, Ottosen C, Soerensen J, Ottesen B. Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale. BJOG 2008;115:908–916.

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.*[3,4] and the Observational Clinical Human Reliability Assessment system[5–7] (OCHRA) by Tang *et al.* The Time-Error matrix has later successfully been modified to a more detailed version,[8] 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.*[9] and Reznick *et al.*[10] 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.[10]

It is well documented that GRSs are reliable, have high interrater reliability (IRR) and construct validity.[9] Since first presented, the OSATS has been modified and tested in many different surgical areas such as open surgery,[11] laparoscopic surgery,[12] vascular surgery[13] and microsurgery,[14] urology,[15] ophthalmology (Global Rating Assessment of Skills in Intraocular Surgery),[16] gynaecology[17,18] and obstetrics.[19,20] 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 0.71[9] to 0.97.[21] 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,[22] 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.[22] In those studies where the IRR is stated, it varies from 0.70[9] to 0.97.[23] 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.[24] 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[25] 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[9] and the modified rating scale for laparoscopic cholecystectomy, developed by Grantcharov *et al.*[12] It consists of five general items and five task-specific items equivalent to the OCRS by Dath *et al.*[24] 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.*[26] (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

Table 1. OSA-LS: assessment chart

	1	2	3	4	5
OSA-LS general skills					
Economy of movements	Many unnecessary movements		Efficient motion but some unnecessary movements		Maximum economy of movements
Confidence of movements: instrument handling	Repeatedly makes tentative or awkward moves with instruments		Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Economy of time	Too long time used to perform sufficiently		Intermediate time used to perform sufficiently		Minimal time used to perform sufficiently
Errors: respect for tissue	Frequently used unnecessary force on tissue or risk of damage by inappropriate use of instruments or instruments often out of sight		Careful handling of tissue but occasionally risk of (minimal) damage, or instruments out of sight		Consistently, handled tissues appropriately with no risk of damage, instruments always in sight
Flow of operation/ operative technique	Imprecise, wrong technique in approaching the operative interventions, or constant supervisor corrections		Careful technique with occasional errors or little supervisor correction		Fluent, secure and correct technique in all stages of the operative procedure, no supervisor corrections
OSA-LS specific skills					
Presentation of anatomic structures	Poor retraction and exposure of fallopian tube and round ligament		Satisfactory retraction and exposure of fallopian tube and round ligament		Expert retraction and exposure of fallopian tube and round ligament
Use of diathermy	Using diathermy too close to healthy ovarian or other tissue, risk of damage		Mostly safe use minimal risk of damage		Perfectly safe use of diathermy, no risk of damage
Dissection of fallopian tube	Inadequate dissection of fallopian tube. Additional damage or bleeding or part of fallopian tube left *in situ*		Identified fallopian tube, adequate dissection little damage of other structures, little bleeding		Clearly identified fallopian tube, perfectly dissected no additional damage, no bleeding. Fallopian tube completely removed
Care for ovary, ovarian artery and pelvic wall	Using diathermy or cutting too close to ovarian artery high risk of bleeding or occlusion of vessel or cauterising ovary		Mostly safe use of instruments, low risk of arterial damage, little cauterising on ovary		Perfectly safe use instruments, no risk of cauterising or cutting the ovary, ovarian artery or other non-target tissue
Extraction of fallopian tube	Clumsily performed with major difficulty to catch the tissue, retract or get the tissue in the bag		Minor difficulty retracting or getting the tissue in the bag		Perfect retraction grasps end of structure or, easy placement of tube in bag

Σ General skills: ———
Σ Specific skills: ———
Σ All skills: ———
Total time in minutes: ———
Non-assessed items: ———

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

Table 2. Operating instruction: laparoscopic salpingectomy, modified after Nezhat *et al.* (2000)[26]

Salpingectomy: operation instruction (expert consensus)	
Instruments	Graspers, bipolar diathermy, scissors, rinse/suction, bag
Procedure	After introducing the trocars and pneumoperitoneum
1	Start the video recording
2	Insert your instruments, grasper in lateral trocar other instrument in medial trocar
3	Identify the anatomy
4	Operate from centre towards lateral
5	Use grasper in right hand and grasp the fallopian tube
6	Use bipolar grasper in left hand and use diathermy on salpinx and mesosalpinx
7	Start close to tubal corner of the uterus
8	Shift bipolar grasper to scissors in left trocar and cut the coagulated tissue close to fallopian tube
9	Continue alternated use of bipolar grasper and scissors to remove the fallopian tube. Use instruments in the trocars providing the most appropriate access to the tissue
10	Take care not to use diathermy on the ovary and the supplying artery and other non-target tissue
11	Use bag or gasper to remove the dissected tissue
12	Use rinse/suction device to clean up blood, use bipolar grasper to coagulate any remaining bleeding vessels/tissue
13	Pull out both instruments and tell the supervisor if you consider the operation performed. Stop video recording

are calculated. Analysis was performed using the Statistical Package for Social Sciences (SPSS®) 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 experienced 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).

The gamma correlation coefficient was 0.91 (95% CI 0.785–1.000) for all observations (Figure 2). The lowest correlation was found in the novice group, the highest correlation in the expert group. Even in the novice group, where the lowest gamma correlation coefficient was found, the discrepancy of the observers' ratings was randomly distributed. This emphasises that none of the observers systematically rated the performances differently from the other observer, i.e. neither more negative nor more positive.

The kappa value on items level (all 21 subjects) varied form 0.510 to 0.933, indicating that item 2, 4 and 6 were main sources of disagreement; in the other items, observers reached a higher degree of agreement (Table 5). The median time used for evaluating the unedited video recordings, including filling out the score table, was 16 minutes (range 7–35). Not all participating gynaecologists used an Endobag® (LiNA Medical UK Ltd, Dulford, UK) or other bag systems to remove the dissected tissue from the body. Some of the women underwent further surgery, e.g. hysterectomy, consequently the fallopian tube was removed en bloc at a later stage, together with additional dissected tissue. Item 10 has therefore been excluded from this validation study.

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

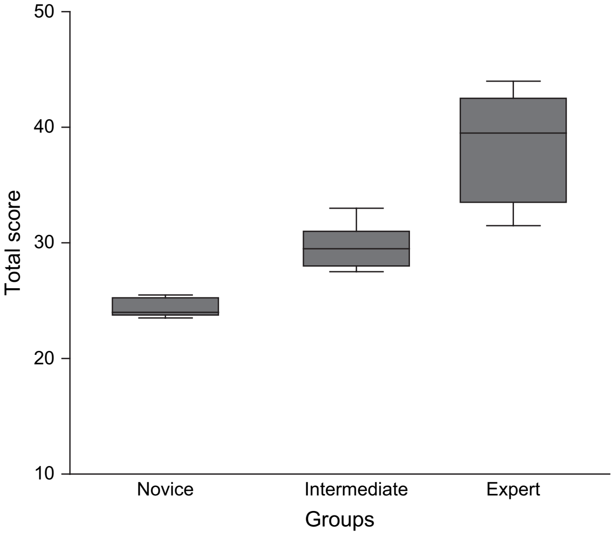

Figure 1. Box plot median score all groups, band represents median; boxes represents IQR; whiskers represents range.

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)

	n	Median	IQR		Range		*P* value	
			25% percentile	75% percentile	Maximum	Minimum	Kruskall–Wallis	Mann–Whitney (*post hoc*)
Novice	8	24.00	23.75	25.25	25.50	23.50	<0.001	Novice vs intermediate <0.03
Intermediate	7	29.50	28.00	31.00	33.00	28.00		Intermediate vs experts <0.03
Expert	6	39.50	33.50	42.50	44.00	31.50		
All	21	30.50	24.50	34.25	44.00	23.50		—

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]

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

Group	*n*	Items evaluated	Numbers of observations	Number of disagreement	IRR
Novice	8	9	72 (9 × 8)	14	0.806
Intermediate	7	9	63 (9 × 7)	6	0.905
Expert	6	9	54 (9 × 6)	13	0.759
Overall	21	9	189 (9 × 21)	32	**0.831**

Item 10 excluded, see text for details.

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

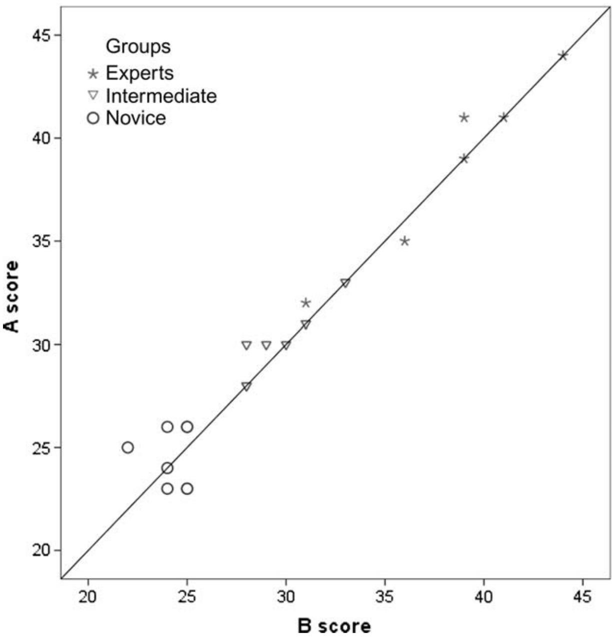

Figure 2. Scatter plot: observer B cumulated score for each individual (*x* axis) versus observer A cumulated score for same subjects (*y* axis). Identification line represents absolute agreement among observers. Distance from line represents magnitude of observer disagreement.

Table 5. Kappa values on single items level

Item	*n*	Kappa	SE	95% CI	
				Lower bound	Upper bound
1	21	0.851	0.100	0.655	1.000
2	21	0.653	0.132	0.394	0.912
3	21	0.933	0.065	0.806	1.000
4	21	0.625	0.129	0.372	0.878
5	21	0.789	0.112	0.569	1.000
6	21	0.510	0.149	0.218	0.802
7	21	0.731	0.147	0.443	1.000
8	21	0.695	0.137	0.426	0.964
9	21	0.695	0.131	0.438	0.952
10	0	—	—	—	—

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.[29] 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. They should, more accurately, until their true technical competence level is established objectively, be called 'quantitative experts' or 'experienced'. As a consequence of this, the educational system is currently replacing training and assessment systems that only counted the number of procedures performed by training and assessment systems based on competence levels.[30] Nevertheless, in most previous studies, experts have only been defined by number of performed procedures, thus leaving this problem as a challenge for future research.

The wider range in performance in the expert group can also be explained by a wider variety of cases operated by the expert group, a *case mix* situation. The novices will always get the easiest cases, and the experts the most complicated cases, such as cases with dense adhesions or significant pathological anatomy. This could influence the assessment. In the expert group, there was one low outlier, which was shown to be a more complicated case, due to a severe hydrosalpinx with dense adhesions to a cystic ovary. This is a weakness of all assessment systems developed from bench station tests, which were originally designed to standardised cases, because they do not take into account that the same procedure in some cases is in reality more difficult in others. There are two ways to overcome this problem. Either the assessment system should only be used for cases of predefined complexity, for instance using only simple and uncomplicated cases for assessment purposes. Another way of dealing with the biological variation is by developing a graded system for case complexity. Multiplying the total score by a predefined factor according to case complexity might solve the problem. This, 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,[31,32] 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.[9]

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.[33] 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,[31,32] 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 items 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 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

1 Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003;327:1032–7.

2 Lidegaard O, Hammerum MS. [The National Patient Registry as a tool for continuous production and quality control]. *Ugeskr Laeger* 2002; 164:4420–3.

3 Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Andersen DK, Satava RM. Analysis of errors in laparoscopic surgical procedures. *Surg Endosc* 2004;18:592–5.

4 Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, *et al*. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg* 2002;236:458–63.

5 Tang B, Hanna GB, Cuschieri A. Analysis of errors enacted by surgical trainees during skills training courses. *Surgery* 2005;138:14–20.

6 Tang B, Hanna GB, Joice P, Cuschieri A. Identification and categorization of technical errors by Observational Clinical Human Reliability Assessment (OCHRA) during laparoscopic cholecystectomy. *Arch Surg* 2004;139:1215–20.

7 Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A. Competence assessment of laparoscopic operative and cognitive skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World J Surg* 2006; 30:527–34.

8 Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA III, *et al*. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg* 2007;193:797–804.

9 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, *et al*. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;84:273–8.

10 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;173:226–30.

11 Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *Am J Surg* 2006;192:372–8.

12 Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg* 2004;91:146–50.

13 Beard JD, Choksy S, Khan S; Vascular Society of Great Britain and Ireland. Assessment of operative competence during carotid endarterectomy. *Br J Surg* 2007;94:726–30.

14 Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS, *et al*. The educational impact of bench model fidelity on the acquisition of technical skill: the use of clinically relevant outcome measures. *Ann Surg* 2004;240:374–81.

15 Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol* 2002;167:1243–7.

16 Cremers SL, Lora AN, Ferrufino-Ponce ZK. Global Rating Assessment of Skills in Intraocular Surgery (GRASIS). *Ophthalmology* 2005;112: 1655–60.

17 Lentz GM, Mandel LS, Goff BA. A six-year study of surgical teaching and skills evaluation for obstetric/gynecologic residents in porcine and inanimate surgical models. *Am J Obstet Gynecol* 2005;193: 2056–61.

18 Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol* 2000;96:146–50.

19 Siddighi S, Kleeman SD, Baggish MS, Rooney CM, Pauls RN, Karram MM. Effects of an educational workshop on performance of fourth-degree perineal laceration repair. *Obstet Gynecol* 2007;109:289–94.

20 Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol* 2006;195:617–21.

21 VanBlaricom AL, Goff BA, Chinn M, Icasiano MM, Nielsen P, Mandel L. A new curriculum for hysteroscopy training as demonstrated by an objective structured assessment of technical skills (OSATS). *Am J Obstet Gynecol* 2005;193:1856–65.

22 Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003;17:1525–9.

23 Goff B, Mandel L, Lentz G, Vanblaricom A, Oelschlager AM, Lee D, *et al*. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol* 2005;192:1331–8.

24 Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, *et al*. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc* 2004;18:1800–4.

25 McLeod PJ, Steinert Y, Trudel J, Gottesman R. Seven principles for teaching procedural and technical skills. *Acad Med* 2001;76:1080.

26 Nezhat C, Siegler A, Nezhat F, Nezhat C, Seidman D, Luciano A. Operations on the Fallopian Tube. In: *Operative Gynecologic Laparoscopy: Principles and Techniques*. San Francisco, CA: McGraw-Hill; 2000. pp. 246–51.

27 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65 (9 Suppl):S63–7.

28 Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc* 2007;21: 2240–3.

29 Grantcharov TP, Funch Jensen P. Learning curve patterns in technical skills acquisition in laparoscopic surgery. Can everybody learn it? Association of Surgeons of Great Britain and England Clinical Congress 3rd to 5th May 2006, Abstract no. 10745.

30 Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001;357:945–9.

31 Larsen CR, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, *et al*. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surg Endosc* 2006;20:1460–6.

32 Aggarwal R, Tully A, Grantcharov T, Larsen CR, Miskry T, Farthing A, *et al*. Virtual reality simulation training can improve technical skills during laparoscopic salpingectomy for ectopic pregnancy. *BJOG* 2006; 113:1382–7.

33 Agresti A. *Analysis of Ordinal Categorical Data*. Hoboken, NJ, USA: Wiley, 1984.