ORIGINAL REPORTS

Validity and Reliability of Global Operative Assessment of Laparoscopic Skills (GOALS) in Novice Trainees Performing a Laparoscopic Cholecystectomy

Kelvin H. Kramp, MD,[*] Marc J. van Det, MD, PhD,[,†] Christiaan Hoff, MD,[*] Bas Lamme, MD, PhD,[‡] Nic J.G.M. Veeger, MSc,[§,‖] and Jean-Pierre E.N. Pierie, MD, PhD[*,¶]*

[*]Department of Surgery, Medical Center Leeuwarden, Leeuwarden, The Netherlands; [†]Department of Surgery, Hospital Group Twente, Almelo, The Netherlands; [‡]Department of Surgery, Albert Schweitzer Hospital, Dordrecht, The Netherlands; [§]Department of Epidemiology, Medical Center Leeuwarden, Leeuwarden, The Netherlands; [‖]Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands; and [¶]Post Graduate School of Medicine, University Medical Center Groningen, Groningen, The Netherlands

PURPOSE: Global Operative Assessment of Laparoscopic Skills (GOALS) assessment has been designed to evaluate skills in laparoscopic surgery. A longitudinal blinded study of randomized video fragments was conducted to estimate the validity and reliability of GOALS in novice trainees.

METHODS: In total, 10 trainees each performed 6 consecutive laparoscopic cholecystectomies. Sixty procedures were recorded on video. Video fragments of (1) opening of the peritoneum; (2) dissection of Calot's triangle and achievement of critical view of safety; and (3) dissection of the gallbladder from the liver bed were blinded, randomized, and rated by 2 consultant surgeons using GOALS. Also, a grade was given for overall competence. The correlation of GOALS with live observation Objective Structured Assessment of Technical Skills (OSATS) scores was calculated. Construct validity was estimated using the Friedman 2-way analysis of variance by ranks and the Wilcoxon signed-rank test. The interrater reliability was calculated using the absolute and consistency agreement 2-way random-effects model intraclass correlation coefficient.

RESULTS: A high correlation was found between mean GOALS score ($r = 0.879$, $p = 0.021$) and mean OSATS score. The GOALS score increased significantly across the 6 procedures ($p = 0.002$). The trainees performed significantly better on their sixth when compared with their first cholecystectomy ($p = 0.004$). The consistency agreement interrater reliability was 0.37 for the mean GOALS score ($p = 0.002$) and 0.55 for overall competence ($p < 0.001$) of the 3 video fragments.

CONCLUSION: The validity observed in this randomized blinded longitudinal study supports the existing evidence that GOALS is a valid tool for assessment of novice trainees. A relatively low reliability was found in this study. (J Surg ■:■■■-■■■. © 2014 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: laparoscopy, trainee, assessment, videotape recording, laparoscopic cholecystectomy

COMPETENCIES: Practice-Based Learning and Improvement, Interpersonal and Communication Skills, Systems-Based Practice

INTRODUCTION

Objective assessment of technical skills of surgical trainees is an important topic in the field of surgical education. To provide a valid and reliable tool in the assessment of surgical skills, Martin et al.[1] developed a global rating scale in the late 1990s, currently known as the Objective Structured Assessment of Technical Skills (OSATS). OSATS has been implemented in many academic centers to measure operative performance in the operating theater and provide feedback to trainees. Although the OSATS is considered to be a validated tool for global assessment of operative competence,[2,3] there was no equivalent for

Correspondence: Inquiries to Kelvin H. Kramp, MD, Department of Surgery, Medical Center Leeuwarden, P.O. Box 888, 8901 BR Leeuwarden, The Netherlands; e-mail: k.h.kramp@gmail.com

  1931-7204/$30.00
http://dx.doi.org/10.1016/j.jsurg.2014.08.006

laparoscopic surgery. As laparoscopic surgery is the standard for an increasing list of procedures, there was a need for a valid and reliable assessment tool that addresses the specific requirements of laparoscopic surgery. Laparoscopic surgery involves a man-machine environment that requires the ability to work with a 2-dimensional view, decreased degrees of freedom, and reduced tactile feedback. Furthermore, the surgeon is challenged by the fulcrum effect, and inversion and scaling of movements of the parts of the instruments inside the abdomen. To evaluate these skills, Vassiliou et al.[4] developed Global Operative Assessment of Laparoscopic Skills (GOALS), a non–procedure-specific assessment tool that can be applied to any procedure in MIS (minimally invasive surgery).

Rasmussens' model of human behavior can be used to describe different levels that have to be achieved in laparoscopic skill training to obtain competency in MIS.[5] In the first level, the trainee acquires skill-based behavior by learning automated sensory-motor patterns. It has been shown that these skills can be improved on a virtual-reality simulator.[6] In the early post-simulator development phase, learned sensory-motor patterns are calibrated to the MIS environment, whereas rule- and knowledge-based behaviors are acquired. Moore and Bennett[7] demonstrated that the risk of complications is approximately 1.7% in the first laparoscopic cholecystectomy and decreases to 0.7% after 5 cases. Although much has changed in the education of trainees, this novice development stage can still be considered as one of the most important learning phases in guiding surgical trainees to competency in performing a laparoscopic cholecystectomy. This study was conducted to explore the validity and reliability of using GOALS for video assessment of laparoscopic cholecystectomy in this critical learning phase.

METHOD

Participants and Patient Selection

In total, 10 surgical residents in their first ($n = 4$) and second ($n = 6$) year of training were recruited for a training curriculum in laparoscopic cholecystectomy. Only trainees who had attended less than 5 laparoscopic procedures and had no experience with performing a laparoscopic cholecystectomy were included. A minimum of 6 months' experience with open surgery was a prerequisite to participate in the study. After a basic laparoscopic skills training, the trainees performed 6 laparoscopic cholecystectomies in the operating room under the supervision of 1 of the 3 participating experienced laparoscopic surgeons.

All the patients included in the study had uncomplicated symptomatic gallstone disease. All the patients gave informed consent before undergoing surgery.

Basic Laparoscopic Skills Training

Basic laparoscopic skills were acquired on the SIMENDO laparoscopy trainer (SIMENDO, Rotterdam, The Netherlands). The intention of the SIMENDO simulator training is to teach trainees a specified level of basic automated sensory-motor patterns required for safe participation in laparoscopic procedures in humans.

Direct Observation: OSATS Assessment

The OSATS was developed by Martin et al.[1] in 1997 and is currently the standard method for the assessment of surgical skills. The OSATS consists of 7 items: (1) respect for tissue, (2) time and motion, (3) instrument handling, (4) knowledge of instruments, (5) use of assistants, (6) flow of operation, and (7) knowledge of the procedure. Each item was scored as generally used in the Dutch surgical training program on a 10-point scale.

The 3 supervising surgeons that randomly supervised the operations used the OSATS to assess the laparoscopic performance of the trainees. Because OSATS assessment is an integral part of the surgical curriculum in the Netherlands, the surgeons had used the OSATS frequently in the past to assess trainees. Although no formal OSATS instruction course was taken, the principles were discussed in teach-the-teacher trainings. The surgeons were uninformed about the number of procedures the trainee performed previously, but not blinded to the identity of the trainee.

To determine whether the increase in OSATS is mainly caused by non–sensory-motor skill acquisition, the OSATS-sensory motor (OSATS-sm) was calculated by summing the items 1, 2, 3, and 6 of the OSATS form.

Video Assessment: GOALS and Overall Competence

The GOALS assessment form contains 6 items. Four items represent domains of technical competence in laparoscopic surgery: (1) depth perception, (2) bimanual dexterity, (3) efficiency, and (4) tissue handling. The fifth item is used to rate the autonomy of the subject. Only parts of the video in which the trainee performed as operating surgeon were edited so the item autonomy was therefore left out of the GOALS form. The sixth item, level of difficulty, was added by Chang et al.[8] to also take into account any difference in difficulty of the procedure.

To be able to compare GOALS with the modified 10-point version of the OSATS global rating scale that is used in our institution, the items on the GOALS form were converted to a 10-point scale. Complementary to the GOALS items, a grade for overall competence was rated on a 10-point scale for each video fragment. It has been shown that transformation of a 5-point scale to a 10-point scale does not significantly influence the data characteristics

besides a slight decrease in the scores with respect to the maximum achievable score.[9]

During every procedure, a video was recorded using the laparoscopic camera, and audio was recorded using 2 microphones: one attached to the trainee and one to the supervising surgeon. The videos were divided into the following 3 sections: (1) opening of the peritoneum, (2) dissection of Calot's triangle and achievement of critical view of safety (CVS), and (3) dissection of the gallbladder from the liver bed. The audio material was used to identify the sections in which the trainee was acting as the operating surgeon. When a supervising surgeon took over the procedure, that part was cut from the video. The video fragments were terminated after 5 minutes or when a section was completed. Subsequently, the videos were muted so the raters were blinded from the performing trainee and the supervising surgeon. After editing and removal of the audio, the order for video assessments was randomized on the basis of participating trainee and number of cholecystectomies performed, whereas the order of the video fragments was maintained. Each individual video fragment was rated by 2 consultant surgeons who were involved in the training program for laparoscopic surgery (Fig. 1).

Statistical Analysis

The usefulness of a measurement tool is dependent on the degree that it measures what it is supposed to measure (validity) and the accuracy of those measurements (reliability). The GOALS scores were used to calculate construct validity (increase in performance score with increase in caseload), concurrent validity (correlation with the OSATS), and interrater reliability (absolute (AA) and consistency agreement (CA) between the 2 raters). SPSS 20.0.0.1 (SPSS, Chicago, IL) was used in all the analyses. Statistical significance was defined as $p < 0.05$.

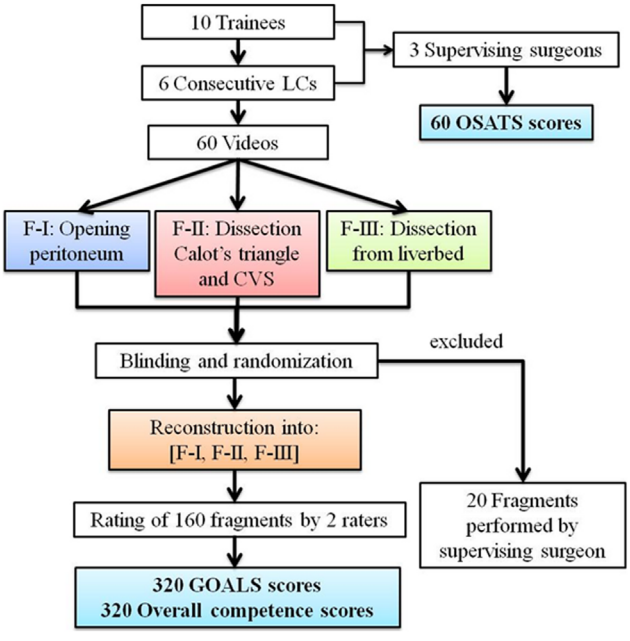

FIGURE 1. Work flow. F, video fragment; LC, laparoscopic cholecystectomy.

Validity

To estimate concurrent validity, the correlation between mean GOALS score and OSATS score of the supervising surgeon was calculated using the Pearson r correlation coefficient. The Friedman 2-way analysis of variance by ranks was used to estimate the construct validity. In addition, the performance on the first was compared with the performance on the sixth cholecystectomy using the Wilcoxon signed-rank test.

Reliability

The intraclass correlation coefficient (ICC) was used to calculate the reliability. Because the ability to estimate progression is the most important aspect of the learning trajectory of the trainees in our study sample, we were interested in the commonly used AA, but also in the CA interrater reliability between both the raters. Therefore, the AA 2-way random-effects model for single measures (AA-ICC 2,1) and the CA 2-way random-effects model for single measures (CA-ICC 2,1) of the ICC were chosen.[10-12]

The mean total GOALS score, the mean items score, and the mean overall competence score of 3 video fragments were compared between both the raters. Values used for ordinal classification of the interrater reliability are always arbitrary in nature and should be adjusted to the purpose of the measurement instrument. Because GOALS would primarily be used for formative assessment in our study population and not for high-stakes examination, cutoff points for the ICC were chosen as "moderate" (0.21-0.40), "reasonable" (0.41-0.60), "good" (0.61-0.80) and "almost perfect" (0.81-1.00).[13]

RESULTS

Measurements

In total, 60 laparoscopic cholecystectomies were successfully recorded. Overall, 160 video fragments were blinded, randomized, and rated by 2 raters. Owing to the intervention of the supervising surgeon, 20 video fragments could not be rated. There were no technical problems. The total yield was 320 measurements (Fig. 1).

As presented in Table 1, the mean OSATS score of the 2 raters was 20.2 ± 8.5 at procedure 1 and increased to 43.5 ± 6.6 at procedure 6. The mean OSATS-sm increased from 10.5 ± 4.1 at procedure 1 to 23.6 ± 4.2 at procedure 6. As presented in Table 2, the mean GOALS score of the 2 raters increased from 20.0 ± 4.8 at procedure 1 to 23.7 ± 4.3 at procedure 6. As presented in Table 2, the mean overall competence score of the 2 raters was 4.6 ± 1.1 at procedure 1 and 5.4 ± 1.0 at procedure 6.

TABLE 1. Mean OSATS Score and Mean OSATS-sm Score (Items 1, 2, 3, and 6 From OSATS) per Caseload

Procedure	1	2	3	4	5	6	p	p (Δ1-6)
OSATS	20.2 ± 8.5	27.5 ± 7.3	34.2 ± 10.0	34.9 ± 11.3	37.6 ± 6.0	43.5 ± 6.6	<0.001*	0.008*
OSATS-sm	10.5 ± 4.1	14.4 ± 3.5	17.8 ± 5.7	18.2 ± 5.7	19.7 ± 4.8	23.6 ± 4.2		

*Statistically significant.

Validity

A high correlation between mean GOALS score and mean OSATS score was observed ($r = 0.879$, $p = 0.021$).

The OSATS scores increased significantly with the caseload ($p < 0.001$), and there was a significant difference between the OSATS scores of the trainees measured in the first vs sixth operation ($p = 0.008$) (Fig. 2). Approximately, 50% of the total increase in OSATS scores consisted of sensory-motor items (Table 1).

The GOALS scores increased significantly with caseload ($p = 0.002$), and there was a significant difference between the GOALS scores of the trainees measured in the first vs sixth operation ($p = 0.004$) (Table 2). The overall competence also increased significantly with the caseload ($p = 0.016$) and between the first and sixth operation ($p = 0.003$) (Table 2).

The GOALS scores and overall competence scores of the video fragments only showed a significant increase in the video fragment of the dissection of Calot's triangle and achievement of CVS ($p = 0.011$ and 0.030, respectively) (Table 2).

Reliability

Table 3 shows the AA-ICC and CA-ICC of the mean total GOALS score, the mean GOALS items score, and the mean overall competence of the 3 video fragments.

The AA-ICC and CA-ICC of the mean GOALS score were moderate (0.37, $p = 0.002$; 0.37, $p = 0.002$) (Fig. 3). The highest AA-ICC was found for the item "efficiency" (0.47, $p < 0.001$) and the lowest for the item "level of difficulty" (0.22, $p < 0.001$). The highest CA-ICC was found for the item "level of difficulty" (0.55, $p < 0.001$) and lowest for the item "bimanual dexterity" (0.27, $p = 0.019$).

The mean overall competence score AA-ICC was moderate (0.36, $p < 0.001$) and the CA-ICC mean overall competence score was reasonable (0.55, $p < 0.001$) (Fig. 4).

DISCUSSION

Objective assessment of surgical trainees is necessary to ensure that professional standards are being met

TABLE 2. Number of Fragments, Mean GOALS Score per Fragment, Mean GOALS Score for 3 Fragments, and Mean Overall Competence Score per Caseload

Procedure	1	2	3	4	5	6	p	p (Δ1-6)
Number of fragments								
N total = 160	21	29	27	27	29	27		
GOALS score								
F1: Opening of the peritoneum	18.9 ± 5.1	20.9 ± 3.9	20.3 ± 4.6	19.3 ± 3.9	24.6 ± 5.6	23.7 ± 5.1	0.063	0.208
F2: Dissection of Calot's triangle and achievement of CVS	19.4 ± 4.9	22.7 ± 4.1	22.3 ± 3.8	25.3 ± 4.3	22.5 ± 3.9	23.9 ± 4.2	0.005*	0.011*
F3: Dissection from the liver bed	21.1 ± 4.6	24.0 ± 4.4	22.6 ± 3.4	24.6 ± 3.4	22.7 ± 3.0	23.6 ± 3.7	0.129	0.447
Mean (F1, F2 and F3)	20.0 ± 4.8	22.6 ± 4.3	21.8 ± 4.0	22.8 ± 4.7	23.2 ± 4.3	23.7 ± 4.3	0.002*	0.004*
Overall competence								
F1: Opening of the peritoneum	4.5 ± 1.2	4.7 ± 1.2	4.7 ± 1.2	4.1 ± 1.3	5.2 ± 1.6	5.2 ± 1.2	0.525	0.305
F2: Dissection of Calot's triangle and achievement of CVS	4.4 ± 1.2	5.3 ± 1.0	5.0 ± 1.0	5.4 ± 1.3	4.9 ± 1.1	5.3 ± 1.0	0.021*	0.030*
F3: Dissection from the liver bed	4.8 ± 1.0	5.4 ± 0.9	5.3 ± 0.9	5.7 ± 1.0	5.1 ± 1.1	5.7 ± 0.7	0.113	0.068
Mean (F1, F2 and F3)	4.6 ± 1.0	5.2 ± 0.8	5.0 ± 0.6	4.9 ± 1.1	5.1 ± 0.9	5.3 ± 0.7	0.016*	0.003*

p Values are based upon the Friedman 2-way analysis of variance by ranks and the Wilcoxon signed-rank test.
*Statistically significant.

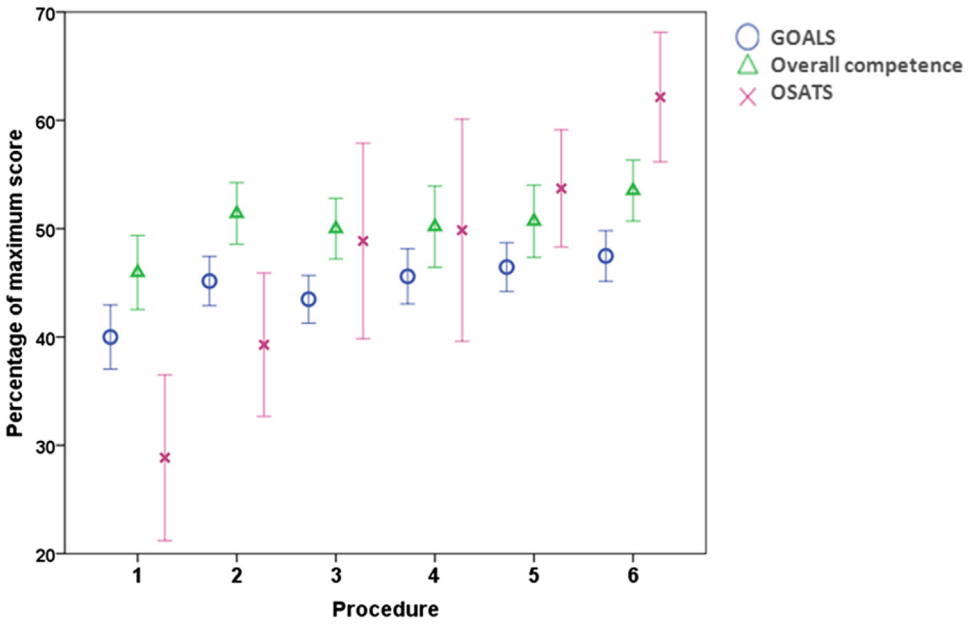

FIGURE 2. Increase in OSATS score, GOALS score, and overall competence score. The difference in OSATS score, GOALS score, and overall competence score in the 6 consecutive procedures was significant ($p = 0.008$, 0.002, and 0.016). Error bars indicate 95% CIs.

in the operating room. In contrast to OSATS, GOALS contains specific criteria for MIS. In this longitudinal study, GOALS was used to assess blinded randomized video fragments of a laparoscopic cholecystectomy.

In total, 6 consecutive cholecystectomies performed by 10 trainees were recorded. Of the video recordings, 3 fragments were edited to produce 160 video fragments, which were assessed by 2 blinded raters previously unexposed to GOALS. The significant increase in mean GOALS score across the 6 procedures and the high correlation with the mean OSATS score indicate that GOALS is a valid instrument for assessment of laparoscopic surgical skills.

Earlier studies have shown that GOALS can distinguish surgeons of varying skill level,[4,8,14,15] but there are only 2 blinded studies that evaluated GOALS.[8,15] The first study did not use repeated measurements of the same trainees, leaving room for individual differences between the trainees to influence the results.[15] The second study was a blinded study that used 2 videos: one of a novice and one of an expert. Although the study was blinded, the high difference in skill performance of the 2 videos was derivable from the video duration time (55 vs 15 min).[8] Both studies indicate that the assessors can thus distinguish a novice from an expert using GOALS, but provide no longitudinal information about the learning curve measured using GOALS. In our study, the increase in performance was tracked with repeated measurements of an identical group of trainees with no prior in vivo laparoscopic experience to highlight the value of GOALS assessment and its implementation in surgical training programs. Furthermore, the video fragments were not only blinded, but also randomized. Raters were therefore not only unaware of the identity of the trainee, but also of the number of cholecystectomies performed previously.

Although the results indicate statistically significant construct validity, the difference in mean GOALS score between the first and sixth procedure was minimal (7%). This may be caused by several factors. Firstly, the high score of approximately 40% of the maximum score in the first procedure suggests that the raters should have been encouraged to make better use of the full range of the items on the GOALS form. Secondly, it could be caused by a "real" high level of sensory-motor skill level in the first procedure achieved through simulator training in the basic laparoscopic training course, although the mean percentage of the maximum score in the first procedure of the OSATS (29%) and OSATS-sm (26%) does not support this. Thirdly, there was an absence of feedback to the trainee based on the GOALS items, as was done using OSATS. Feedback gives the trainee the opportunity to strengthen his or her weaknesses and achieve a higher score in the assessment of a subsequent performance.

TABLE 3. ICC of the Mean GOALS Score of the Items, the Mean Total GOALS Score, and Overall Competence

Item	Domain	AA-ICC (2,1)	CA-ICC (2,1)
1	Depth perception	0.23	0.33
2	Bimanual dexterity	0.26	0.27
3	Efficiency	0.47	0.47
4	Tissue handling	0.34	0.33
5	Level of difficulty	0.22	0.55
Overal competence		0.36	0.55
GOALS score		0.37	0.37

All ICC-values were statistically significant ($p < 0.05$).

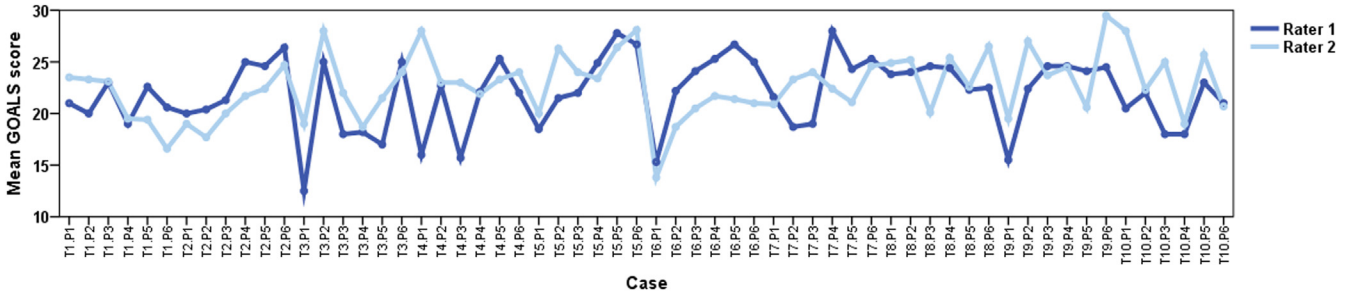

FIGURE 3. Interrater reliability of mean GOALS score of fragment 1 to 3 between rater 1 and 2. T, trainee; P, procedure.

FIGURE 4. Interrater reliability of mean overall competence score of fragment 1 to 3 between rater 1 and 2. T, trainee; P, procedure.

In this study, a significant increase in mean GOALS score and mean overall competence score was only observed in the video fragment of the dissection of Calot's triangle and achievement of CVS. These results are consistent with those observed by Aggarwal et al. using motion tracking data.[16] Aggarwal et al. found a statistically significant difference in time taken, total path length, and number of movements in the video fragment of the dissection of Calot's triangle between a novice and an experienced group. They did not find any difference using motion tracking data in clipping and cutting of the cystic artery, in clipping and cutting of the cystic duct, and in the dissection of the gallbladder from the liver bed in path length and number of movements. The most likely explanation for these observations is that the dissection of the Calot's triangle is the hardest step to complete. As a result, it is probably the most sensitive procedural step for operative performance measurements such as GOALS assessment, overall competence scores, or motion tracking data.

A low interrater reliability was observed in the mean of the 3 video fragments of one procedure performed by a trainee (0.37). This ICC means that 37% of the difference between ratings is attributable to true variance and the remaining variance is attributable to random error, rater error, and/or other sources of error.[17]

There are multiple factors that can influence the reliability in assessments. An important factor is the training of the raters in the assessment method. The lowest reliability of GOALS was reported by Vassiliou et al.[15] In this study, Vassiliou et al. compared direct observation ratings with blinded videotape ratings. They found an ICC of 0.39 when the scores of one of the video raters were compared with those of 2 direct raters. They ascribe the ICC of 0.39 to the video raters' lack of previous exposure to GOALS. But Vassiliou et al. also describe a video rater that was in like manner unexposed to GOALS, but attained an ICC of 0.76 when his scores were compared with the 2 direct observations. This video rater reported to have invested a considerable amount of time in getting comfortable with the assessment method by watching all the videos beforehand and watching videos multiple times during the assessment.[15] According to the authors, these findings suggest that training in GOALS assessment might be necessary before reliable GOALS scores can be achieved. Matsuda et al. had similar findings in their study of the Endoscopic Surgical Skill Qualification System in Japan; the amount of exposure to their assessment method correlated significantly with the reliability of the ratings. They stated that long-term experience with their assessment method may be necessary to be able to perform reliable skill assessments.[18] The results of this study might likewise indicate that the interrater reliability is jeopardized when GOALS is used without proper instructions and training of the raters.

A second contributing factor lies in the calculation used for estimating the ICC. The ICC harbors the variance within the sample to calculate the reliability. As the estimated true variance on the basis of the heterogeneity within the sample decreases, the calculated ICC automatically tends to decrease.[17] The CA was unexpectedly higher for overall competence (0.55) than for the total GOALS score (0.37), while an equally low or lower ICC for overall

competence is expected when the calculated true variance in the sample is too small to achieve an acceptable ICC. Furthermore, the ICC model that is used for calculating reliability is often not reported. The results may vary significantly when different calculation models for the ICC are used.[10-12,17]

A third explanation could be in the scale used in the GOALS form. Some authors state that attaining an AA reliability of 0.80 is one of the major inherent difficulties of using a Likert scale.[19]

Finally, although raters involved in surgical education are probably inclined to invest energy and time in the assessment, their motivation may be threatened by mental fatigue or time pressure and therefore lead to unreliable measurements.[20] Practical consequences of this may be that video assessments are limited to a particular section of the operation or raters are rewarded to guarantee sufficient motivation.

Limitations of this Study

Although our measurements indicate that GOALS has significant construct and concurrent validity when assessing novice trainees, some limitations should be kept in mind. Firstly, different methods were used for OSATS and GOALS assessment; OSATS assessment was performed with direct observation and GOALS assessment with video fragments. Secondly, we used 320 GOALS assessments to measure the improvement in surgical skills. Our large sample size probably disguised the low interrater reliability and made it possible to establish validity. Therefore, the question remains whether the validity also exists in the operating theater when the measurement of skill level is based on one rating. However, the validity could be higher because the rater is not blinded, the item autonomy is included, and the rater assesses the whole procedure instead of only fragments. Thirdly, assessing a consecutive series of 6 identical procedures probably seldom takes place during a residency. In most cases, there is an interval of learning without a formal assessment of the trainee. Finally, although the raters were consultant surgeons who were familiar with assessing trainees, we did not identify whether there existed a difference in the perception of what can be defined as "good" or "bad" laparoscopic skills.

In the field of MIS, there is a demand for objective assessment of professional skills to meet increasing political and public demands. The availability of an objective assessment method gives educators the opportunity to certify trainees according to their abilities. The certification enables a formal, transparent, and objective identification of trainees who are able to complete laparoscopic procedures independently, skillfully, and most importantly, safely. The OSATS can be considered as an option, but the results of recent studies have raised concerns about the objectivity of the OSATS and some authors therefore reject the idea that OSATS can function as an instrument for summative assessment.[3,21] GOALS could be an alternative to the OSATS, but it is important to mention that the reliability found in our study sample cannot be generalized to trainees in higher ranges of surgical skill level.

CONCLUSIONS

In conclusion, this randomized, blinded, longitudinal study supports the existing evidence that GOALS has construct and concurrent validity for the assessment of novice trainees performing a laparoscopic cholecystectomy. The reliability observed in this study was relatively low.

REFERENCES

1. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278.
2. Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today*. 2013;43(3):271-275.
3. Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. *Can J Surg*. 2011;54(2):116-122.
4. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005;190(1):107-113.
5. Wentink M, Stassen LP, Alwayn I, Hosman RJ, Stassen HG. Rasmussen's model of human behavior in laparoscopy training. *Surg Endosc*. 2003;17(8):1241-1246.
6. Ikonen TS, Antikainen T, Silvennoinen M, Isojärvi J, Mäkinen E, Scheinin TM. Virtual reality simulator training of laparoscopic cholecystectomies—a systematic review. *Scand J Surg*. 2012;101(1):5-12.
7. Moore MJ, Bennett CL. The learning curve for laparoscopic cholecystectomy. The Southern Surgeons Club. *Am J Surg*. 1995;170(1):55-59.
8. Chang L, Hogle NJ, Moore BB, et al. Reliable assessment of laparoscopic performance in the operating room using videotape analysis. *Surg Innov*. 2007;14(2):122-126.
9. Dawes John. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales *Int J Market Res*. 2008;50(1):61-77.

10. McGraw KO, Wong SP. Forming inferences about some intra-class correlation coefficients. *Psychol Methods*. 1996;1(1):30-46.

11. Shrout PE, Fleis JL. Intra-class correlation: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2): 420-428.

12. Krebs DE. Declare your ICC type. *Phys Ther*. 1986; 66(9):1431.

13. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1): 159-174.

14. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007;204(2):308-313.

15. Vassiliou MC, Feldman LS, Fraser SA, et al. Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. *Surg Innov*. 2007;14(3):211-216.

16. Aggarwal R, Grantcharov T, Moorthy K, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. 2007;245(6):992-999.

17. Portney LG, Watkins MP. *Foundations of Clinical Research—Application to Practice*, 3rd ed. 2007; New Jersey: Pearson Education, Inc, 07458.

18. Matsuda T, Kanayama H, Ono Y, et al. Reliability of laparoscopic skills assessment on video: 8-year results of the endoscopic surgical skill qualification system in Japan. *J Endourol*. 2014 [Epub ahead of print].

19. 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(10):1525-1529.

20. Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci Biobehav Rev*. 2014;44C:58-75.

21. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br J Surg*. 2010;97(7): 972-987.