Education in laparoscopic surgery
Kramp, Kelvin Harvey

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 7

VALIDITY AND RELIABILITY OF GLOBAL OPERATIVE ASSESSMENT OF LAPAROSCOPIC SKILLS (GOALS) IN NOVICE TRAINEES PERFORMING A LAPAROSCOPIC CHOLECYSTECTOMY


Journal of Surgical Education 2015;72(2):351-358
Abstract

**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:** Ten 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 (CVS) and 3) dissection of the gallbladder from the liver bed were blinded, randomized and rated by two consultant surgeons with GOALS. Also, a grade was given for overall competence. The correlation of GOALS with live observation OSATS scores (Objective Structured Assessment of Technical Skills) was calculated. Construct validity was estimated with the Friedman two-way analysis of variance by ranks and Wilcoxon signed-rank test. The inter-rater reliability was calculated with the absolute and consistency agreement two-way random effects model intra-class 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 ($p = 0.004$). Consistency agreement inter-rater reliability was 0.37 for the mean GOALS score ($p = 0.002$) and 0.55 for overall competence ($p = 0.002$) 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. Compared to other studies a low reliability was found in this study.
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. developed a global rating scale in the late 1990s, currently known as the OSATS (Objective Structured Assessment of Technical Skills). OSATS has been implemented in many academic centers to measure operative performances in the operating theater and provide feedback to the trainee. Although the OSATS is considered to be a validated tool for global assessment of operative competence, there was no equivalent for laparoscopic surgery. Since 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; inversion and scaling of movements of the parts of the instruments inside the abdomen. To evaluate these skills Vassiliou et al. developed GOALS (Global Operative Assessment of Laparoscopic Skills), a non-procedure specific assessment tool that can be applied to any procedure in minimally invasive surgery (MIS).

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. 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. In the early post-simulator development phase, learned sensory-motor patterns are calibrated to the MIS environment while rule- and knowledge-based behaviors are acquired. Moore & Bennet 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 since then, 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
Ten 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 OR under the supervision of one of the three participating surgeons. All patients included in the study had uncomplicated symptomatic gallstone disease. All 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. 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 three 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. 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-sm (OSATS-sensory-motor) 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 5th 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 6th item, level of difficulty, was added by Chang et al. 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 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 slightly decrease in the scores with respect to the maximum achievable score.

During every procedure a video was recorded with the laparoscopic camera and audio was recorded with 2 microphones; one attached to the trainee and one to the supervising surgeon. The videos were divided into 3 sections: 1) opening of peritoneum, 2) dissection of Calot’s Triangle and achievement of 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 for the performing trainee and the supervising surgeon. After editing and removal of the audio, the order for video assessments was randomly set on the basis of participating trainee and number of cholecystectomies performed while the order of the video fragments was maintained. Each individual video fragment was rated by two consultant surgeons who were involved in the training program for laparoscopic surgery (Figure 1).

![Figure 1: Workflow. F=Video fragment; LC=laparoscopic cholecystectomy](image)

**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 and consistency agreement between two raters). SPSS 20.0.0.1 (SPSS, Chicago, IL, USA) was used in all analysis. Statistical significance was defined as \( p < 0.05 \).

**Validity**

To estimate concurrent validity, the correlation between mean GOALS score and OSATS score of the supervising surgeon was calculated with the Pearson’s \( r \) correlation coefficient. The Friedman two-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 with the Wilcoxon signed-rank test.

**Reliability**

The intra-class 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 absolute agreement, but also in the
consistency agreement inter-rater reliability between the two raters. Therefore, the absolute agreement two-way random effects model for single measures (AA-ICC 2,1) and the consistency agreement two-way mixed effects model for single measures (CA-ICC 3,1) of the intra-class correlation coefficient were chosen.\textsuperscript{10-12}

The mean total GOALS score, the mean items score and mean overall competence score of 3 video fragments was compared between the two raters. Values used for ordinal classification of the inter-rater 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, cut-off points for the ICC were chosen as ‘moderate’ (0.21 to 0.40), ‘reasonable’ (0.41 to 0.60), ‘good’ (0.61 to 0.80) and ‘almost perfect’ (0.81 to 1.00).\textsuperscript{13,14}
Results

Measurements
Sixty laparoscopic cholecystectomies were successfully recorded. A total of 160 video fragments were blinded, randomized and rated by 2 raters. Twenty video fragments could not be rated due to intervention of the supervising surgeon. There were no technical problems. The yield was 320 measurements (Figure 1).

As presented in table 1, the mean OSATS score of the two 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. The mean GOALS score of the two raters increased from 20.0±4.8 at procedure 1 to 23.7±4.3 at procedure 6. The mean overall competence score of the two 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 (item 1, 2, 3 and 6 from OSATS) per caseload.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>p</th>
<th>p(Δ1-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSATS-sm</td>
<td>10.5±4.1</td>
<td>14.4±3.5</td>
<td>17.8±5.7</td>
<td>18.2±5.7</td>
<td>19.7±4.8</td>
<td>23.6±4.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSATS</td>
<td>20.2±8.5</td>
<td>27.5±7.3</td>
<td>34.2±10.0</td>
<td>34.9±11.3</td>
<td>37.6±6.0</td>
<td>43.5±6.6</td>
<td>&lt;0.001*</td>
<td>0.008*</td>
</tr>
</tbody>
</table>

* Statistical 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 caseload (p < 0.001) and there was a significant difference between the OSATS scores of the trainees measured in the first versus sixth operation (p = 0.008) (Figure 2). Approximately 50% of the total increase in OSATS scores consisted of sensory-motor items (Table 1).

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, p = 0.002 and p = 0.016). Error bars indicate 95% confidence intervals.

The GOALS scores increased significantly with caseload (p < 0.001) and there was a significant difference between the GOALS scores of the trainees measured in the first versus sixth operation (p =
The overall competence also increased significantly with 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 (Table 2).

Table 2: Number of fragments, mean GOALS score per fragment, mean GOALS score for 3 fragments and mean overall competence score per caseload.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>p</th>
<th>p(Δ1-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N total=160</td>
<td>21</td>
<td>29</td>
<td>27</td>
<td>27</td>
<td>29</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOALS score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1: Opening of the peritoneum</td>
<td>18.9±5.1</td>
<td>20.9±3.9</td>
<td>20.3±4.6</td>
<td>19.3±3.9</td>
<td>24.6±5.6</td>
<td>23.7±5.1</td>
<td>0.063</td>
<td>0.208</td>
</tr>
<tr>
<td>F2: Dissection of Calot’s Triangle and achievement of CVS</td>
<td>19.4±4.9</td>
<td>22.7±4.1</td>
<td>22.3±3.8</td>
<td>25.3±4.3</td>
<td>22.5±3.9</td>
<td>23.9±4.2</td>
<td>0.005*</td>
<td>0.011*</td>
</tr>
<tr>
<td>F3: Dissection from the liver bed</td>
<td>21.1±4.6</td>
<td>24.0±4.4</td>
<td>22.6±3.4</td>
<td>24.6±3.4</td>
<td>22.7±3.0</td>
<td>23.6±3.7</td>
<td>0.129</td>
<td>0.447</td>
</tr>
<tr>
<td>Overall competence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1: Opening of the peritoneum</td>
<td>4.5±1.2</td>
<td>4.7±1.2</td>
<td>4.7±1.2</td>
<td>4.1±1.3</td>
<td>5.2±1.6</td>
<td>5.2±1.2</td>
<td>0.525</td>
<td>0.305</td>
</tr>
<tr>
<td>F2: Dissection of Calot’s Triangle and achievement of CVS</td>
<td>4.4±1.2</td>
<td>5.3±1.0</td>
<td>5.0±1.0</td>
<td>5.4±1.3</td>
<td>4.9±1.1</td>
<td>5.3±1.0</td>
<td>0.021*</td>
<td>0.030*</td>
</tr>
<tr>
<td>F3: Dissection from the liver bed</td>
<td>4.8±1.0</td>
<td>5.4±0.9</td>
<td>5.3±0.9</td>
<td>5.7±1.0</td>
<td>5.1±1.1</td>
<td>5.7±0.7</td>
<td>0.113</td>
<td>0.068</td>
</tr>
</tbody>
</table>

P-values are based upon the Friedman two-way analysis of variance by ranks and the Wilcoxon signed-rank test.* Statistical significant

Reliability

Table 3 shows the AA-ICC and CA-ICC of the mean total GOALS score, the mean GOALS items score and mean overall competence of the 3 video fragments. The AA-ICC and CA-ICC of the mean GOALS score were moderate \((0.37; \ p = 0.002, 0.37; \ p = 0.002)\) (Figure 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’ \((0.27; \ p = 0.019)\).
Figure 3: Inter-rater reliability of mean GOALS score of fragment 1 to 3 between rater 1 and 2.

Figure 4: Inter-rater reliability of mean overall competence score of fragment 1 to 3 between rater 1 and 2.

Table 3: The AA- and CA-ICC of the mean GOALS score of the items, the mean total GOALS score and overall competence score of the assessment of three video fragments of a procedure of one trainee (N=320).

<table>
<thead>
<tr>
<th>Item</th>
<th>Domain</th>
<th>AA-ICC (2,1)</th>
<th>CA-ICC (3,1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Depth perception</td>
<td>0.23</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>Bimanual dexterity</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>3</td>
<td>Efficiency</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>4</td>
<td>Tissue handling</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>5</td>
<td>Level of difficulty</td>
<td>0.22</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>Overall competence score</td>
<td>0.36</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>GOALS score</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

All ICC-values were statistical significant (p < 0.05)
Objective assessment of surgical trainees is necessary to ensure professional standards are being met in the operating room. In contrast to OSATS, GOALS contains specific criteria for minimally invasive surgery. In this longitudinal study GOALS was used to assess blinded randomized video fragments of a laparoscopic cholecystectomy.

Six consecutive cholecystectomies performed by 10 trainees were recorded. Out of the video recordings 3 fragments were edited to produce a total of 160 video fragments which were assessed by two 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.

Validity
Earlier studies have shown that GOALS can distinguish surgeons of varying skill level, but there are only two blinded studies that evaluated GOALS. The first study did not use repeated measurements of the same trainees, leaving room for individual differences between trainees to influence the results. 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 two videos was derivable from the video duration time (55 min. vs. 15 min.). Both studies indicate that assessors can thus distinguish a novice from an expert with GOALS, but provide no longitudinal information about the learning curve measured with 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 statistical 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. First, the high score of approximately 40% of the maximum score in the first procedure suggests that the raters should have been encouraged more to use the full range of the items on the GOALS form. Second, 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%) do not support this. A third possible cause is the absence of feedback to the trainee based on the GOALS items, as was done with the 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.

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. with motion tracking data. Aggarwal et al. found a statistical 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 experienced group. They did not find any difference with motion tracking data in clipping and cutting of the cystic artery, 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.
Reliability

A low reliability was observed in the mean of the three video fragments of one procedure performed by a trainee (0.37). The low ICC means that a low percentage (37%) of the difference between ratings is attributable to true variance and the remaining variance is attributable to other sources.

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. 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. Vassiliou et al. also describe a video rater that was in like manner unexposed to the 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. 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 is necessary to perform reliable skill assessments. The results of this study might indicate likewise that the interrater reliability is jeopardized when GOALS is used without proper instructions and/or 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 between-subject variance decreases the calculated ICC automatically tends to decrease.

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

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. 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. First, different methods were used for OSATS and GOALS assessment; OSATS assessment was performed with direct observation and GOALS assessment with video fragments. Second, we used a total of 320 GOALS assessments to measure the improvement in surgical skills. Our large sample size probably disguised the low inter-rater 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 a smaller sample of measurements. On the other hand, the validity could be higher because the item autonomy is included and/or the rater assesses the whole procedure instead of only fragments. Third, 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 formal assessment of the trainee. Fourth, although the raters were consultant surgeons that 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 minimally invasive surgery, there is a demand for objective assessment of professional skills in order 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. Certification enables a formal, transparent and objective identification of trainees who are able to complete laparoscopic procedures independently, skillfully and most important, 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.\textsuperscript{20,21} GOALS could be a better alternative to the OSATS for this purpose. Therefore, 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 assessment of novice trainees performing a laparoscopic cholecystectomy. The reliability observed in this study was low compared to the reliability found in other studies.
References