Value of an objective assessment tool in the operating room

Ellen Hiemstra, MD *
Wendela Kolkman, MD, PhD*
Ron Wolterbeek, MSc†
Baptist Trimbos, MD, PhD*†
Frank Willem Jansen, MD, PhD *‡

From the Departments of *Gynecology and †Medical Statistics and Bioinformation, Leiden University Medical Center, Leiden, and the ‡Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, the Netherlands

Accepted for publication Mar. 31, 2010

Correspondence to:
Dr. E. Hiemstra
Department of Gynecology
Leiden University Medical Center
P.O. Box 9600, K6-76
2300 RC Leiden
The Netherlands
E.Hiemstra@lumc.nl

DOI: 10.1503/cjs.032909

Background: Concerns about the achievement of surgical proficiency during residency are increasing. To objectify surgical skills, the Objective Structured Assessment of Technical Skills (OSATS) was developed and proven valid, feasible and reliable for use in laboratory settings. This study aimed to evaluate the value of this tool for intraoperative use.

Methods: Residents were assessed with an OSATS after every procedure they performed as the primary surgeon during a 3-month clinical rotation in gynecological surgery. We mapped individual learning curves (OSATS scores plotted against experience) and established the average procedure-specific learning curve. We used linear mixed models to assess the relation between performance and experience.

Results: Nine residents were recruited and 319 OSATS analyzed. Individual learning curves revealed progression beyond 24 of 30 OSATS points for 7 residents. Performance on the average procedure improved with experience, and the OSATS score increased by an average of 1.10 points per assessed procedure (\( p = 0.008, 95\% \text{ confidence interval } 0.44–1.77\)). Median OSATS scores ranged from 18 to 30 among the 21 assessors.

Conclusion: Intraoperative implementation of OSATS seems to offer important advantages: structured feedback is facilitated, and learning curves enable insight into individual progression. However, doubts have been raised about the objectivity of the tool. Therefore, caution is warranted in using it for graduation and certification.

Contexte : La maîtrise de la chirurgie acquise au cours de la résidence préoccupe de plus en plus. Pour rendre les techniques chirurgicales objectives, on a mis au point l’évaluation structurée objective des habiletés techniques (Objective Structured Assessment of Technical Skills; OSATS) et on en a démontré la viabilité, la faisabilité et la fiabilité pour utilisation en laboratoire. Cette étude vise à déterminer la valeur de cet outil au cours des interventions.

Méthodes : Les médecins résidents ont été soumis à une évaluation OSATS après chaque intervention qu’ils ont pratiquée comme chirurgien principal au cours d’un stage clinique de 3 mois en chirurgie gynécologique. Nous avons établi les courbes d’apprentissage individuelles (courbes des résultats OSATS en fonction de l’expérience), ainsi que la courbe d’apprentissage moyenne particulière à l’intervention. Nous avons utilisé des modèles mixtes linéaires pour évaluer le lien entre le rendement et l’expérience.

Résultats : Nous avons recruté 9 médecins résidents et analysé 319 évaluations OSATS. Les courbes d’apprentissage individuelles ont révélé un progrès supérieur à 24 points OSATS sur 30 chez 7 médecins résidents. L’exécution de l’intervention moyenne s’est améliorée avec l’expérience et le résultat de l’évaluation OSATS a augmenté en moyenne de 1,10 point par intervention évaluée (\( p = 0.008, \text{ intervalle de confiance à } 95 \%, 0,44–1,77\)). Les résultats médians de l’évaluation OSATS ont varié de 18 à 30 chez les 21 évaluateurs.

Conclusion : L’application de l’évaluation OSATS aux interventions semble offrir les avantages importants : les commentaires structurés sont facilités et les courbes d’apprentissage donnent une idée de l’évolution de chaque médecin. On a toutefois soulèvé des doutes quant à l’objectivité de l’outil. La prudence s’impose donc dans l’utilisation de celui-ci aux fins d’obtention du diplôme et du certificat.

Nowadays, it is becoming more and more difficult to achieve surgical proficiency. Residents experience less training owing to reduced working hours and a decreased surgical caseload. Additionally, with the development of new surgical techniques, skills acquisition is more challenging. Currently, basic surgical procedures are sufficiently mastered after finishing
residency training, but advanced procedures are not. Ultimately, skills deficiencies will impede postresidency performance. Moreover, residency programs still rely heavily on informal and subjective evaluations based on recollections of supervisors. Therefore, on one hand, surgical skills training needs to become more efficient, and on the other hand, appropriate assessment is required to optimally benefit from the spare learning moments in the operating room (OR).

An objective assessment tool can fulfill an important role during operative training. Such a tool can help the learning process through constructive feedback on performance. Second, an assessment tool can be applied to establish competency levels and to mark progression. Finally, it can provide benchmark criteria to be used as a training goal or for credentialing purposes.

To fulfill this need for an objective assessment tool, the Objective Structured Assessment of Technical Skills (OSATS) was developed by Martin and colleagues in Toronto in 1997. An OSATS consists of a procedurespecific checklist, a pass/fail judgment and a global rating scale. The latter turned out to be superior in terms of reliability and validity. On this global rating scale, domains are scored on a Likert scale ranging from 1 to 5, with an explicit description at points 1, 3 and 5.

So far, studies about the quality of OSATS have mainly been conducted in simulators or live animal models. Although applying OSATS in simulator settings has the benefit of repeated practice without the risk of harming patients, simulators will never perfectly mimic operative conditions. Therefore, OSATS have been implemented for the assessment of real surgical procedures on a large scale in residency programs in the Netherlands. Moreover, plans are being developed to use this form of assessment tool for certification purposes after residency training. However, only a few studies have investigated the value of the intraoperative use of OSATS.

Aggarwal and colleagues found that the OSATS score discriminates between a novice and an expert surgeon performing a laparoscopic cholecystectomy demonstrated by video-based assessment. Bodle and colleagues concluded from feedback questionnaires that trainers and trainees in the United Kingdom perceived the OSATS to be valid and valuable. In the absence of data on the implementation of OSATS in daily practice, the current study was conducted to assess the value of the tool in clinical practice by analyzing residents' learning curves for a variety of surgical procedures in gynecology.

METHODS

In the Netherlands, the obstetrics and gynecology (Ob/Gyn) residency program lasts 6 years. On average, 3 of these 6 years are spent in a university teaching hospital, and the complementary period is spent in a nonuniversity teaching hospital. The university hospitals provide a curriculum to train residents in a variety of subspecialties, like reproductive health care, perinatology and oncology. Specifically, a 3-month clinical rotation is spent on gynecological surgery. During this rotation, which is generally attended during the fourth postgraduate year (PGY-4), residents are scheduled to perform surgery in the OR 4 days a week. Gradually, a resident is given more responsibility as experience accrues, depending on the resident's technical skills, the type of procedure and patient characteristics. Finally, a resident performs a procedure as the primary surgeon in the presence of a supervising consultant.

Study design

In 2005, the global rating scale of the OSATS (referred to as “OSATS” in this paper) was introduced at the department of Ob/Gyn of the Leiden University Medical Center in an observational study of its implementation in clinical practice (Box 1). The assessment tool had been adapted from Martin and colleagues. The 6 domains of an OSATS represent aspects of technical competence in surgery. The only modification to the original form is that we merged the domains “knowledge of instruments” and “instrument handling.” This is in accordance with the version of the OSATS form used by the Royal College of Obstetrics and Gynecology. During this implementation study, residents were instructed to register an OSATS assessment of every procedure that they performed as a primary surgeon during their 3-month rotation in gynecological surgery. Procedures during which a resident independently performed some important steps were included as well. After the supervising consultant had filled out the OSATS form, the results were discussed with the resident to provide him/her with constructive feedback per domain.

Whereas the assessed trainees were PGY-4 Ob/Gyn residents, the assessors could be any gynecologist working as a consultant in the department who was supervising the surgical procedure. They were instructed on how to complete the OSATS form. In essence, the instruction was to mark the number on the Likert scale corresponding to the resident’s performance on each domain, irrespective of the training level.

Individual learning curves

All OSATS were collected, and data were analyzed using SPSS version 16.0 (SPSS Inc.). The total score of each OSATS was calculated by adding up the score of the 6 domains (with a minimum possible score of 6 and a maximal possible score of 30 points). An OSATS score of 24 points equals the score in which each domain on average is rated with 4 points (75% of the maximal score that ranges from 1 to 5). This score was chosen as a threshold for good surgical performance in the absence of benchmark criteria in other studies. Learning curves for each individual resident were drawn by plotting his or her...
OSATS scores against the total caseload during a clinical rotation, regardless of which procedures were performed. To establish the caseload, all consecutively performed procedures that were assessed with an OSATS were numbered. For each resident, the mean OSATS score during the rotation was calculated, and progression in time was illustrated by mapping a regression line.

Construct validity

No “gold standard” is available to measure surgical performance. Therefore, the construct validity (i.e., the extent to which a test measures the trait that it purports to measure) should be used to verify the quality of an assessment tool for surgical skills. In this study, the construct validity of OSATS was established by testing the hypothesis that surgical performance improves as the procedure-specific experience accrues. For that purpose, the average learning curve for the “average” procedure was mapped by plotting the OSATS score against the procedure-specific caseload. The procedure-specific caseload was also based on the number of assessed procedures.

To test this hypothesis, a linear relation between OSATS score and experience was assumed. The advantage of simplifying the average procedure-specific learning curve to a straight line is that the performance level at the start can be determined, as well as the amount of progression in technical surgical skills, taking individual performance levels and learning potential into account. Therefore, a linear mixed model was fitted as a random coefficients model with a random slope and intercept per resident.

We considered \( p < 0.05 \) to be statistically significant, and we calculated 95 percent confidence intervals (CIs).

Objectivity of assessment with OSATS

After this implementation study, we sought the opinions of assessed trainees and assessors regarding the objectivity of assessment with an OSATS. They were asked to rate the OSATS on a Likert scale ranging from 1 “subjective” to 5 “objective.” The assessed trainees were residents who were recruited during an education afternoon in the Leiden University Medical Center, of which the attendance was obligatory during Ob/Gyn residency training. The assessors were the same consultants who had participated in the implementation study.

RESULTS

Nine residents attended a 3-month clinical rotation in gynecological surgery and agreed to participate in the study; 3 men and 6 women. Nineteen different types of procedures were assessed with an OSATS, and the total number of procedures assessed was 319. Among these procedures, 39% were abdominal, 31% were laparoscopic, 20% were procedures with a vaginal approach and the

<table>
<thead>
<tr>
<th>Box 1. Objective Structured Assessment of Technical Skills (OSATS): global rating scale of operative performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please circle the number corresponding to the candidate’s performance in each category, irrespective of training level</td>
</tr>
<tr>
<td>Respect for tissue:</td>
</tr>
<tr>
<td>1 Frequently used unnecessary force on tissue or caused damage</td>
</tr>
<tr>
<td>2 Careful handling of tissue but occasionally caused inadvertent damage</td>
</tr>
<tr>
<td>3 Consistently handled tissues appropriately with minimal damage</td>
</tr>
<tr>
<td>Time and motion:</td>
</tr>
<tr>
<td>1 Many unnecessary moves</td>
</tr>
<tr>
<td>2 Efficient time/motion but some unnecessary moves</td>
</tr>
<tr>
<td>3 Clear economy of movement and maximum efficiency</td>
</tr>
<tr>
<td>Knowledge and handling of instrument:</td>
</tr>
<tr>
<td>1 Lack of knowledge of instruments</td>
</tr>
<tr>
<td>2 Competent use of instruments but occasionally appeared stiff or awkward</td>
</tr>
<tr>
<td>3 Demonstrated some forward planning with reasonable progression of procedure</td>
</tr>
<tr>
<td>4 Obvious familiarity with instruments</td>
</tr>
<tr>
<td>Flow of operation:</td>
</tr>
<tr>
<td>1 Frequently stopped procedure and seemed unsure of next move</td>
</tr>
<tr>
<td>2 Demonstrated some forward planning with reasonable progression of procedure</td>
</tr>
<tr>
<td>3 Obviously planned course of procedure with effortless flow from one movement to the next</td>
</tr>
<tr>
<td>4 5</td>
</tr>
<tr>
<td>Use of assistants:</td>
</tr>
<tr>
<td>1 Consistently placed assistants poorly or failed to use assistants</td>
</tr>
<tr>
<td>2 Appropriate use of assistants most of the time</td>
</tr>
<tr>
<td>3 Strategically used assistants to the best advantage at all times</td>
</tr>
<tr>
<td>4 5</td>
</tr>
<tr>
<td>Knowledge of specific procedure:</td>
</tr>
<tr>
<td>1 Deficient knowledge. Needed specific instructions at most steps</td>
</tr>
<tr>
<td>2 Knew all important steps of procedure</td>
</tr>
<tr>
<td>3 4 5</td>
</tr>
</tbody>
</table>
remaining 10% were hysteroscopies (Table 1). On an individual basis, the median number of procedures assessed was 40 (range 12–60).

**Individual learning curves**

The 9 individual learning curves were drawn by plotting OSATS scores against the total caseload, regardless of which specific procedure had been performed, during the clinical rotation (Fig. 1). The regression lines of these curves are displayed too, together with the threshold of 24 (of a possible 30) OSATS points. Regression analysis revealed that the 2 residents with the lowest average scores (residents A and B) did not reach the threshold of 24 points within their clinical rotations. Residents C and D reached the threshold when nearing the end of their rotations. Residents H and I achieved relatively high scores at the start of the 3-month period and continued to show improvement.

**Average procedure-specific learning curve**

Additionally, we plotted the average OSATS scores against the experience (i.e., the procedure-specific caseload) for a particular procedure passed the threshold of 24 points at a caseload of 5 procedures. Additionally, a plateau in performance was reached after a caseload of 8 procedures. To establish the construct validity of OSATS, we tested whether the OSATS score increased significantly with an increasing caseload using a linear mixed model. The slope of the general learning curve was 1.10 OSATS points per assessed procedure ($p = 0.008$, 95% CI 0.44–1.77). In other words, the average performance based on total OSATS score improved by 1.10 points for every consecutively performed procedure.

**Objectivity of the assessment**

The assessors were 21 gynecologists, all working as consultants at the Department of Gynecology at the Leiden University Medical Center. The median OSATS scores given to residents by each assessor ranged from 18 to 30, and the number of assessed procedures ranged from 1 to 114. Some gynecologists assessed only 1 specific procedure (e.g., a cesarean section), whereas others assessed the entire surgical spectrum.

All 24 residents who were present at the obligatory education afternoon answered the question about the objectivity of assessment with OSATS. One person who was just beginning residency was excluded from analysis owing to inexperience with this assessment form. Residents rated the OSATS with a median score of 2 (range 1–4) on a 5-point Likert scale, with 1 indicating “subjective” and 5 indicating “objective.” The median score of the assessors was 3 (range 1–4).

**DISCUSSION**

Intraoperative OSATS can be used to assess residents’ surgical training over time. By plotting the OSATS score against experience, it can be determined whether and how much progression has occurred. The use of an objective assessment tool is a new way to establish learning curves. Prior parameters are the duration of surgery, the complication rate and the conversion rate in the case of laparoscopic procedures. However, duration of surgery and complication rate have been shown to be crude and indirect, as these indicators largely depend on the difficulty of the individual surgical case (e.g., the comorbidity of a patient) and the supervising surgeon. The intraoperative use of OSATS may overcome these disadvantages.

Two of 9 residents did not progress beyond the benchmark level of 24 of 30 OSATS points within the 3-month clinical rotation. This failure is likely to be a sign of stagnation of their learning processes and can only partially be explained by the coincidence that they encountered more complex procedures later in their rotations than the other residents. Additionally, only 2 residents showed good performance, taking the average OSATS scores and the progression into account, during the entire clinical rotation. This small proportion illustrates the concern about
whether current residency programs with work hour restrictions are sufficient to master surgical proficiency.

The construct validity of the OSATS for assessment purposes was revealed by confirming the hypothesis (i.e., the construct) that a resident’s OSATS score improves as procedure-specific experience accrues. This is not the conventional way to prove the construct validity; however, it is a more subtle approach than the often-used method of confirming the ability of an assessment tool to discriminate between 2 groups of hugely varying levels of experience. The latter method was used by Aggarwal and colleagues,7 who revealed that experienced surgeons have higher OSATS scores than novice surgeons for 1 standardized procedure: the laparoscopic cholecystectomy. The straight line model we used as an argument for the construct validity has 2 limitations. Surgical performance cannot infinitely improve (the maximum OSATS score is 30 points), and the learning curve for surgical skills consists of an initial steep phase, then changes slowly until the curve becomes more flat.21 However, the advantage of simplifying a resident’s learning curve to a straight line and additional analysis with linear mixed model, is that progression in surgical skills can be quantified taking the individual level of performance and learning potential into account. From these data, we found that a resident’s performance improves by an average of 1.1 OSATS points every time the same procedure is performed (and assessed). Of course, we may not simply generalize this conclusion because this increase was based on the average of 19 very different surgical procedures.

The previously mentioned formation of a plateau in performance was observed in the average procedure-specific learning curve. This plateau was achieved after a caseload of 8 (of the same) procedures. This was in accordance with the results of a questionnaire administered
among residents in which they deemed that 10 of the same procedures needed to be performed to be a safe and confident surgeon.\(^2\) Again, the value of this generalization is limited because of the heterogeneous range of assessed procedures.

This study was conducted under regular clinical conditions. Therefore, even the same procedures widely varied with respect to difficulty and risk of complication. Also, variation will have been present in the extent to which consultants allowed residents to independently perform a surgical procedure. Furthermore, the assessment rate might not be 100%. The resulting selection bias may be in favour of the best performed procedures. However, not all procedures need to be assessed to gain insight in the progression of an individual resident. More importantly, the intended objectivity of assessment with an OSATS seems to be disappointing, taking into account our finding that none of the residents or staff members rated the OSATS to be objective. Additionally, the number of assessed procedures and the OSATS score varied enormously among the consultants. This variation occurred despite the uniform instruction that all assessors had received. An attempt to achieve more uniformity might be realized by organizing additional training for the assessors in the registration of an OSATS. However, in our opinion, the effect of such training is limited. No information can be added to the original instruction to mark the number on the rating scale corresponding to the resident’s performance on each domain, irrespective of the training level. Moreover, an assessment based on the opinion of an individual will never be free from subjectivity. A study in which residents all perform at least 10 of the same procedures consecutively would have allowed firmer conclusions about the learning curve for that specific procedure. However, insight in daily practice is obtained by analyzing the heterogeneous data of our study and illustrates the study’s relevance.

**CONCLUSION**

Assessment with OSATS during residency has many advantages. Learning curves based on OSATS have the potential to identify residents in need of more guidance. Consequently, cues are provided to tailor surgical skills training to individual needs. An OSATS does not need to concern the entire procedure; individual steps of the procedure can be evaluated as well. Additionally, an OSATS provides a framework of structured instantaneous feedback on surgical skills in general (total OSATS score). Theoretically, the specific domains of technical skills (e.g., respect for tissue, knowledge and handling of instruments) also provide cues for identifying individual needs. However, the information that the domain-specific scores add is limited, as revealed by the small variety of scores within 1 OSATS. Ideally, the structural feedback on surgical performance using assessment with OSATS will enhance the efficiency of the spare learning moments in the OR. From that point of view, we consider the general global rating scale of OSATS to be suitable for large-scale implementation in the OR.

However, the inherent subjectivity of assessment using an opinion-based tool needs to be taken into account. Regarding the results of the questionnaire and the enormous variation in assessors’ scores, an OSATS unfortunately is not as objective as it intended to be. This is an important limitation of the OSATS that, to our knowledge, has not been highlighted in other studies about this assessment tool. Furthermore, there are other ways to evaluate surgical skills. Therefore, caution needs to be exercised in using OSATS for certification and qualification purposes, or in advising an individual resident to choose a nonsurgical specialization if the OSATS-based performance continues to be disappointing. Though, presently, it seems to be the best tool available.

**Acknowledgements:** The authors thank all residents and consultants for their participation in the study.

**Competing interests:** None declared.

**Contributors:** Dr. Hiemstra wrote the article. All authors helped with study design, acquired and analyzed the data, reviewed the article and approved its publication.
References


Canadian Surgery FORUM

The Canadian Surgery FORUM canadien de chirurgie will hold its annual meeting Sept. 15–18, 2011, in London, Ontario. This interdisciplinary meeting provides an opportunity for surgeons across Canada with shared interests in clinical practice, continuing professional development, research and medical education to meet in a collegial fashion. The scientific program offers material of interest to academic and community surgeons, residents in training and students.

The major sponsoring organizations include the following:
- The Canadian Association of General Surgeons
- The Canadian Society of Colon and Rectal Surgeons
- The Canadian Association of Thoracic Surgeons
- The Canadian Society of Surgical Oncology

Other participating societies include the American College of Surgeons, the Canadian Association of Bariatric Physicians and Surgeons, the Canadian Association of University Surgeons, the Canadian Hepato-Pancreato-Biliary Society, the Canadian Undergraduate Surgical Education Committee, the James IV Association of Surgeons, the Quebec Surgical Association and the Trauma Association of Canada.

For registration and further information contact surgeryforum@rcpsc.edu; www.cags-accg.ca.